

**THE ROLE OF REINTERPRETATION IN THE REVISION OF IMPLICIT  
IMPRESSIONS**

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements of the Degree of

Doctor of Philosophy

by

Thomas Cornell Mann

August 2017

© 2017 Thomas Cornell Mann

# THE ROLE OF REINTERPRETATION IN THE REVISION OF IMPLICIT IMPRESSIONS

Thomas Cornell Mann, Ph.D.

Cornell University 2017

Much research and theory has suggested that implicit (versus explicit) impressions of other people may be particularly difficult to reverse once formed, as they may be relatively insulated from propositional thinking. Furthermore, the revision of *negative* first implicit impressions has been considered particularly difficult to achieve, due to the dominance of negative information in impression formation. I demonstrate across multiple lines of work that negative implicit impressions can be updated when new information provides a reinterpretation of the basis of the initial negative impression, overturning the negative construal of earlier learning.

In a first set of studies (1a-6), people who formed a negative implicit first impression of a new person based on seemingly bad behavior – breaking into and damaging the homes of his neighbors – significantly reversed their implicit impressions after learning that he was actually saving children trapped in a fire. The studies isolate reinterpretation as the mechanism of change through comparison to other positive behaviors that do not reinterpret the earlier story and mediation of the effect through self-reported extent of reinterpretation, and show that revision endures days after the initial study session.

A second line of work (Studies 7-12) examines revision through reinterpretation under a broader set of conditions and domains, to begin to establish the more general applicability and

utility of this mechanism for implicit revision. Studies 7-8 examine the process components of reinterpretation-driven change more closely with the aim of identifying a broader family of strategies for revision. Study 9 assesses whether reinterpretation can still reverse implicit impressions two days after formation of the first impression. Study 10 examines reinterpretation in the domain of public opinion on big-game hunting, while Studies 11 and 12 test the effect of race and gender of the target on the effectiveness of reinterpretation, respectively.

A final line of work examines the automaticity features of reinterpretation-driven updating. Additional analyses of the implicit measures reported in earlier sections show distributional features that have not been reported previously in the literature (bimodality), and then aim to establish the robustness of revision effects to potential explicit influences on the implicit measures and elimination of the unusual distributions. Further studies supported the implicit nature of revision, demonstrating a lack of online participant awareness of their implicit responses (Studies 13-14), persistence of revision effects under taxing distraction (Study 15), and replication of implicit updating with heightened warnings to follow task instructions and elimination of the anomalously bimodal distributions (Studies 16-17).

Collectively, these three sets of findings demonstrate that negative implicit first impressions can be reversed by a reinterpretation of the information upon which a first impression is based, and that this process may contribute to impression revision in a variety of domains.



## **BIOGRAPHICAL SKETCH**

Thomas Mann grew up in Dunstable, Massachusetts. Though surrounded in his hometown by more bovines than hominids, he developed a keen interest in understanding the workings of the human mind, ultimately leading him to complete a degree in Psychology at Tufts University in 2011. He immediately entered graduate school at Cornell University the following Fall. Thomas will further avoid real-world responsibility by next embarking on a two-year National Science Foundation Postdoctoral Fellowship at Harvard University.

## ACKNOWLEDGMENTS

As clichéd as it must be to say, this dissertation truly could never have come to fruition without the help and support of so many mentors, friends, family members, and colleagues. I cannot possibly complete this incredible journey without acknowledging the people who have nurtured my life and career, and without whom I truly could never have reached this point.

I have so many more reasons to be thankful to Melissa Ferguson than I could ever hope to express. Over years of long coffee-fueled walks around Beebe Lake while brainstorming study ideas, summertime lab luncheons at Carriage House, email threads long enough to break Gmail, Skype calls with collaborators, practice talks, pep talks, soul searches, and celebrations, I have found in you not only an amazing advisor, but also a dear friend. I will strive to emulate your wisdom, thoughtfulness, kindness, and steadfast mentorship in everything I do.

I also wish to thank the other members of my committee, Tom Gilovich, Vivian Zayas, and David Smith, for their generous support and encouragement over these last six years. Our discussions have never ceased to be memorable, stimulating, and inspiring, and I am grateful to have had the opportunity to learn from each of you during my time at Cornell. The breadth and depth of my knowledge would be greatly impoverished if not for your mentorship.

I would like to thank my family for all that they have done and sacrificed to make it possible for me to follow my dreams, while reminding me to laugh and not take myself too seriously. I also thank my friends for the web of support that I have felt throughout my life in so many countless ways, big and small. The world filtered through you has been rich and colorful.

Last but not least, I thank Kirk, for being by my side through it all. Your love has touched every part of my life. Thank you for your infinite support, without which I cannot imagine having filled these pages. Your care and encouragement mean the world to me.

## TABLE OF CONTENTS

Biographical Sketch	v
Acknowledgments	vi
Table of Contents	vii
List of Figures	viii
List of Tables	xiii
Chapter I. Introduction	1
Chapter II. Reversal of Novel Implicit First Impressions	11
Chapter III. Mechanisms and Boundary Conditions	103
Chapter IV. Operating Characteristics of Reinterpretation-Based Revision	221
Chapter V. General Discussion	301
Appendix	340

## LIST OF FIGURES

<i>Figure 1.</i> Mean proportion of ideographs judged more pleasant than average in Experiment 1a, by face prime, time, and story condition.....	31
<i>Figure 2.</i> Mean <i>D</i> scores in Experiment 1b, by measurement time and story condition.....	38
<i>Figure 3.</i> Mean proportion of ideographs judged more pleasant than average in Experiment 2, by measurement time, story condition, and face prime.....	45
<i>Figure 4.</i> Mean proportion of ideographs judged more pleasant than average in Experiment 3 at time 2 by cognitive load condition, story condition, and face prime.....	53
<i>Figure 5.</i> Mediation of story condition effect on time 2 judgments of ideograph pleasantness following Francis West primes, through subjective change in meaning of the time 1 information and general extensiveness of thinking in Experiment 5.....	69
<i>Figure 6.</i> Mean proportion of ideographs rated as more pleasant than average in Experiment 6 by measurement time, story condition, and face prime.....	75
<i>Figure 7.</i> Mean explicit liking of Francis West in each story condition at each measurement time in Experiment 6.....	76
<i>Figure 8.</i> Mean proportion of pictographs judged as more pleasant than average in Study 7 by measurement time, story condition, and face prime.....	118
<i>Figure 9.</i> Mean explicit liking of Francis West in Study 7, by story condition and measurement time.....	122
<i>Figure 10.</i> Multinomial process tree for the AMP model in Study 8 for pleasant (+) and unpleasant (-) responses, adapted from Payne et al. (2010).....	126
<i>Figure 11.</i> Mean proportion of pictographs judged as more pleasant than average in Study 8 by measurement time, story condition, and face prime.....	131

<i>Figure 12.</i> Mean proportion of pictographs judged to be more pleasant than the average pictograph, by quiz condition, time, story condition, and person prime, in Study 9.....	153
<i>Figure 13.</i> Example painting stimuli from Study 10.....	171
<i>Figure 14.</i> Mean proportion of paintings judged more pleasant than average, by podcast condition, time, and prime type, Study 10.....	174
<i>Figure 15.</i> The proportion of pictographs judged to be more pleasant than the average pictograph in Study 11, by information condition, Frank’s race, time, and person prime.....	183
<i>Figure 16.</i> The proportion of pictographs judged to be more pleasant than average in Study 12 (PMP condition), by information condition and prime.....	202
<i>Figure 17.</i> Mean <i>D</i> -score index of relative implicit judgment of Jonathan as doctor and Elizabeth as nurse in Study 12 (IAT condition), by information condition and materials condition.....	204
<i>Figure 18.</i> Standardized implicit judgment of Jonathan as more doctor-like (vs. nurse-like) than Elizabeth in Study 12, by information condition, materials, and implicit measure.....	206
<i>Figure 19.</i> Frequency distributions of the proportion of pictographs judged to be more pleasant than average at Time 1 by prime stimulus, pooled across information conditions and experiments (Studies 1a and 2-9).....	224
<i>Figure 20.</i> Values randomly drawn from a normal distribution split into 21 bins, which Hartigan’s dip test registers as multimodal.....	226
<i>Figure 21.</i> Frequency distribution of the proportion of pictographs judged to be more pleasant than average on trials in which Francis West is the prime stimulus at Time 2 in the Fire Rescue condition, pooled across experiments (Studies 1a and 2-9).....	227

<i>Figure 22.</i> Frequency distributions of the proportion of pictographs judged to be more pleasant than average on trials in which Francis West is the prime stimulus at Time 2 in the Fire Rescue condition, Studies 1a and 2-9.....	228
<i>Figure 23.</i> Frequency distributions of the proportion of pictographs judged to be more related to doctors (vs. nurses) than average in Study 12, by prime stimulus and information condition...	229
<i>Figure 24.</i> Frequency distributions of the proportion of pictographs judged to be more pleasant than average at Time 1 in Study 11, by prime stimulus and race of Frank.....	230
<i>Figure 25.</i> Frequency distributions of the proportion of pictographs judged to be more pleasant than average on trials in which Frank West is the prime stimulus at Time 2 in Study 11, by information condition and race of Frank.....	231
<i>Figure 26.</i> Frequency distributions of the proportion of pictographs judged to be more pleasant than average on trials in which Knowlton (the big-game hunter) is the prime stimulus in Study 10, by time and podcast condition.....	232
<i>Figure 27.</i> Predicted AMP results from multinomial model fit to individual participant data, which reproduce the results of Study 8.....	242
<i>Figure 28.</i> Frequency distributions of the predicted proportions of pictographs judged to be more pleasant than average at Time 1 by the AMP model, by prime type.....	242
<i>Figure 29.</i> Frequency distributions of the predicted proportions of pictographs judged to be more pleasant than average at Time 2 by the AMP model, by prime type and information condition.....	243
<i>Figure 30.</i> Frequency distributions of the A parameters fit by the AMP model at Time 1, by Prime Type.....	244

<i>Figure 31.</i> Frequency distributions of the A parameters fit by the AMP model at Time 2, by Prime Type and Information Condition.....	245
<i>Figure 32.</i> Frequency distributions of the M parameters fit by the AMP model, by Information Condition.....	246
<i>Figure 33.</i> Frequency distribution of the proportion of pictographs judged to be more pleasant than average on trials in which Francis West was the prime stimulus in Study 13.....	256
<i>Figure 34.</i> The proportion of pictographs judged to be more related to doctors (over nurses), by prime type and information condition in Study 14.....	262
<i>Figure 35.</i> Frequency distributions of the proportion of pictographs judged to be more related to doctors (vs. nurses) than average in Study 14, by prime type and information condition.....	263
<i>Figure 36.</i> Proportion of pictographs judged more pleasant than average in Study 15, by icon side, icon expression, and prime type.....	273
<i>Figure 37.</i> Frequency distribution of the proportion of pictographs judged to be more pleasant than average on trials in which Francis West was the prime image, in Study 15.....	274
<i>Figure 38.</i> The final page of instructions on the AMP in Studies 16-17.....	281
<i>Figure 39.</i> Mean proportion of paintings judged more pleasant than average in Study 16, by information condition, time, and prime type.....	283
<i>Figure 40.</i> Frequency distribution of the proportion of paintings judged to be more pleasant than average on trials in which Francis West was the prime image at Time 1 in Study 16.....	284
<i>Figure 41.</i> Frequency distributions of the proportion of paintings judged to be more pleasant than average on trials in which Francis West was the prime image in Study 16, by information condition.....	284

<i>Figure 42.</i> Mean explicit evaluations of Francis West in Study 16, by time and information condition.....	287
<i>Figure 43.</i> Mean proportion of paintings judged as more pleasant than average in Study 17, by information condition, time, and prime type.....	292
<i>Figure 44.</i> Frequency distribution of the proportion of paintings judged to be more pleasant than average on trials in which Francis West was the prime image at Time 1 in Study 17.....	293
<i>Figure 45.</i> Frequency distributions of the proportion of paintings judged to be more pleasant than average on trials in which Francis West was the prime image in Study 17, by information condition.....	293



## LIST OF TABLES

<i>Table 1.</i> Correlations between implicit and explicit evaluations and questionnaire measures in fire rescue condition at time 2, experiments 1-6.....	33
<i>Table 2.</i> Zero-order correlations between all variables in the mediation model in Experiment 5.....	68
<i>Table 3.</i> AMP multinomial process model parameter estimates, Study 8.....	135

## **CHAPTER I. Introduction**

For Richard Jewell, the tide of public opinion changed swiftly. On July 27, 1996, tragedy struck the Summer Olympics in Atlanta, Georgia: three pipe bombs detonated inside a backpack in Centennial Olympic Park, killing one bystander and maiming over 100 more. Yet, as often occurs in the face of disaster, an ordinary person donned the hero's mantle. Jewell, a security guard at the park, reportedly discovered the bomb mere minutes before detonation, leaving enough time for him to alert authorities and clear many innocent people out of the area, minimizing loss of life in the process. Jewell was hailed as a hero in the eyes of the nation and the world.

Within days, however, new developments in the investigation cut short the admiration. An anonymous source leaked that Jewell was the prime suspect of the investigation, and the heroic narrative in the media gave way to a more sinister one: Had Jewell planted the bomb himself, perhaps to allow himself to play a hero and win national commendation, or to throw investigators off his trail? Just as swiftly as the nation had sung his praises before, it now turned on him with potent vengeance, assailing his character, excavating his personal history for any tidbits that might fit the growing view of him as a radicalized terrorist, and following the every move of authorities as they interviewed Jewell's friends and acquaintances. It would be 6 years before the saga would come to a close with the apprehension of the true culprit, Eric Rudolph, who carried out the attack in protest of tolerance for abortion in the United States that he viewed as unacceptable – long after the damage to Jewell's reputation had been done. Jewell's story, like so many others, is a testament to the complex problem we all face as we try to understand and evaluate other people – a process that requires a readiness to incorporate new information as it comes to light, and attention to the proper way to construe the meaning of the actions of others.

*Construal* has a well-deserved place among the central concepts of psychology. The idea that objective reality is less psychologically important than how an individual mind interprets and constructs the world is both foundational and enduring, and has deeply informed myriad areas of research, including persuasion (Greenwald, 1968), person impressions and attitudes (Higgins, Rholes, & Jones, 1977; Srull & Wyer, 1979, 1980), judgment and decision making (Vallone, Griffin, Lin, & Ross, 1990), memory (Bartlett, 1932), morality (Graham, Haidt, & Nosek, 2009; Kreps & Monin, 2014; Schein & Gray, in press), emotion (Barrett, 2012; Blechert, Sheppes, Di Tella, Williams, & Gross, 2012; Brooks, 2014; Shurick et al., 2012), intergroup interaction (Yogeeswaran & Dasgupta, 2008), motivation (Fujita & Carnevale, 2012), and more (Trope & Liberman, 2010). The mind is built to construe the world, and construe it does, on scales both macro and micro: The quest to draw meaning from the world consumes years of devoted and concerted work on the part of painters, poets, writers, philosophers, scientists, and graduate students alike, but also is evident in every waking moment as the mind slices and dices raw input to effortlessly categorize, judge, and interact with an environment. The mind is so expert in such tasks that mountains of work now show that countless complex responses can be computed rapidly, with little effort, attention, and even awareness (e.g., Bargh, 1994; Bijleveld, Custers, & Aarts, 2009; Kiesel, Kunde, Pohl, Berner, & Hoffmann, 2009; Marien, Custers, Hassin, & Aarts, 2012; Pessiglione et al., 2009; Rule & Ambady, 2008; Tabak & Zayas, 2012; van Gaal et al., 2014; Willis & Todorov, 2006; see Dehaene, 2014; Hassin, 2013).

That the mind is a construal machine fits hand in glove with the perspective that at multiple levels, it is also a predictive one (Bar, 2011; Clark, 2013; Edelman, 2008; Gregory, 1980; Huang & Rao, 2011; Helmholtz, 1860). To survive and thrive requires efficient computation of the implications of external events and possible courses of action; construal acts

in the service of prediction. It is largely for this reason that evaluation – the basic determination of goodness or badness, or likelihood of helping or harming – has been given such a central role as a fundamental process of cognition (Barrett & Bliss-Moreau, 2009; Cacioppo, Gardner, & Berntson, 1999; Russell, 2003). Given the importance of other people to the life of any individual, it will come as no surprise that much evidence shows that humans are adept at evaluating *social* targets. Inferences about other people are made rapidly and spontaneously, on a variety of dimensions, and even outside of conscious awareness (e.g., Cunningham et al., 2004; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015; Todorov & Uleman, 2002, 2003, 2004; Whalen et al., 1998; Willis & Todorov, 2006).

For all of the importance of accurately drawing actionable inferences about other people, a basic question arises: How well does our inferential machinery actually serve us? How accurate are our immediate responses, in the moment, to other people – responses that prepare us to approach or avoid? And, most pertinent to the present work: How effectively can our responses to others be *updated* in light of new information that we learn over time?

There is much work establishing conditions under which people fail to update erroneous impressions – especially *negative* impressions – of other people. Perceivers may simply avoid stimuli that are viewed as negative, thereby never having an opportunity to encounter disconfirming information, but also minimizing risk of harm for their caution (Eiser, Fazio, Stafford, & Prescott, 2003; Eiser, Stafford, & Fazio, 2008; Fazio, Eiser, & Shook, 2004; Fazio, Pietri, Rocklage, & Shook, 2015). Work on self-fulfilling prophecies has similarly shown that an initial impression can bring about its own confirmation by eliciting corresponding behavior from its target, such as when an expectation that someone will be warm and attractive, or hostile and aggressive, elicits such behavior through one's own actions toward the person (Chen & Bargh,

1997; Snyder, Tanke, & Berscheid, 1977). More generally, confirmation biases are a well-studied quirk of human reasoning that can lead perceivers to selectively seek out or notice information that fits with their expectations (Darley & Fazio, 1980).

Though it is clear that people may fall short of accuracy in their perceptions of others when relevant information is ambiguous, mixed, or risky to attain, one might expect that a mind adept at navigating the complex social world would be *capable* of integrating new information about other people when it *is* available, clear, and seemingly compelling. Yet this expectation begets a puzzle, for decades of research and theory in social-cognitive psychology have crafted a picture of the mind at odds with itself. This research suggests that although many people may be outwardly egalitarian, fair-minded, and quick to incorporate new information in their *explicit* (i.e., intentional) judgments about other people, those same individuals can show stereotype-laden, prejudice-based, intractable responses to others in their *implicit* (i.e., unintentional) responses (Greenwald & Banaji, 1995; Gregg, Seibt, & Banaji, 2006; McConnell, Rydell, Strain, & Mackie, 2008; Nosek, 2007; Nosek, Banaji, & Greenwald, 2002; Rydell & McConnell, 2006; Rydell, McConnell, Strain, Claypool, & Hugenberg, 2007). Stemming largely from dual-systems (Rydell & McConnell, 2006; Smith & DeCoster, 2000) and dual-process (Fazio, 2007; Gawronski & Bodenhausen, 2006, 2011; Petty et al., 2006) accounts of cognition, much of this work has held, and found, that while it is relatively easy to set aside an initial negative impression in explicit judgments, it is more difficult to do so in implicit reactions (McConnell et al., 2008; Rydell & McConnell, 2006; Gawronski, Deutsch, Mbirkou, Seibt, & Strack, 2008). A chief implication of this set of findings and perspective is that people may be incapable of completely integrating compelling, accurate information into their impressions of another person,

leaving their implicit responses yoked to some extent to discredited, rejected information that they have explicitly discarded.

In the present work, I critically reexamine the potential for negative implicit impressions of other people to be rapidly updated in light of new information. The investigation will argue that one potential mechanism through which updating could occur – reinterpretation of earlier learning – has not previously been thoroughly tested as a potentially successful route of implicit impression updating. After presenting a number of studies that provide evidence for the basic efficacy of reinterpretation-driven reversal of initially negative implicit first impressions, I will move toward investigations of boundary conditions, mechanisms, generalizations, and implications of this form of revision, as well as a closer examination of the automaticity features of such updating. In so doing, this work aims to begin to build a picture of a mind that *can* update to reflect a new understanding of the social world, even at an implicit level.

### **Outline of subsequent chapters**

#### ***Chapter II***

In this first central chapter, I present a reproduction of a paper published in 2015 with coauthor Melissa J. Ferguson in the *Journal of Personality and Social Psychology*, titled “Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations.” Across seven studies (1a-6), this paper develops my initial argument that reinterpretation is a largely untested, yet theoretically promising, route through which even initially negative implicit impressions of other people could be durably and quickly reversed. The opening sections of the paper review the substantial body of literature on implicit impression change, introducing the foundational theories and studies that have been taken as evidence for the comparably slow rate of change to which implicit social responses are largely believed to be restricted. The

introductory sections of the paper also review the dual-mode perspectives that have both informed and been informed by the aforementioned work, together with recent theoretical treatments that leave more room open for rapid implicit updating (e.g., De Houwer, 2014).

Turning to the experiments, the initial three studies (Studies 1a, 1b, 2) in the paper aim to establish the basic possibility of rapid reversal of negative implicit first impressions and demonstrate that only new positive information that does provide a reinterpretation of the earlier negative details, and not positive information that fails to do so, results in the reversal of implicit evaluations toward the impression target. These initial studies also introduce the novel theoretical paradigm (the “Francis West” story) that will be featured in most of the studies within this package.

In subsequent studies, Chapter II will further corroborate a reinterpretation-based account of the demonstrated implicit revision effects by showing that a strong cognitive distraction curtails revision even though participants encode the new, reinterpreting information well enough to identify it later (Study 3), and by showing that revision is mediated by self-reported reinterpretation but not by self-reported general extensiveness of deliberation and speed of thinking (Studies 4-5). These studies help to inform both the mechanism and operating conditions (cognitive efficiency) of reinterpretation-driven change. A final study in Chapter II (Study 6) examines the “meaningfulness” of the revision by assessing whether it endures for three days after the end of the first part of the experiment, finding that this revision is durable over that time period, rather than an ephemeral deviation that rapidly reverts to the initial impression.

### ***Chapter III***

After demonstrating the initial revision effect and pinpointing reinterpretation as the critical feature of the new information and cognitive mechanism underlying this reversal, I turn in Chapter III to a deeper investigation of the mechanism, boundary conditions, and generalizability of implicit revision through reinterpretation. The studies in this chapter feature a mix of published and unpublished work.

In Studies 7-8, I aim to more deeply conceptualize and model the constituent processes through which reinterpretation may enact implicit revision by construing reinterpretation as a potential member of a broader family of strategies for change that combine negation of an earlier impression with its simultaneous replacement with a new one, drawing inspiration from various prominent theories of implicit updating as reviewed in Chapter II. Study 7 approaches this analysis experimentally, including multiple conditions meant to isolate these component steps and compare reinterpretation to each, as well as their combination. In the next experiment (Study 8), I supplement the experimental isolation of these processing components with a multinomial process modeling approach designed to examine which constituent processes that occur during the implicit measure itself are affected by reinterpretation, and compare the parameters fit by the model between the different conditions (Batchelder & Riefer, 1999; Conrey et al., 2005; Jacoby, 1991; Payne, 2001; Payne et al., 2010; Sherman et al., 2008).

In Study 9, I present a paper published with coauthor Melissa J. Ferguson in the *Journal of Experimental Social Psychology* in 2017, titled “Reversing implicit first impressions through reinterpretation after a two-day delay.” This study, with its accompanying introduction and discussion, examine the question of whether reinterpretation can be effective in reversing implicit first impressions after a larger time delay (two days) than that used in the previous studies in this package and in most other work in this field (a single lab session). The principal



finding of this work – that reinterpretation was still able to reverse implicit evaluations after the delay, and across varying levels of participant recall for the story details provided in the earlier session – is a critical first test of the broader utility of reinterpretation as a mechanism of implicit change toward real-world targets. The study also affords an opportunity to begin to connect work on implicit impression change with broader trends in the memory literature on how representations can be updated after initial formation with relevant new information (Lee, Nader, & Schiller, 2017).

The subsequent studies in this chapter continue to shift toward examining the generalizability and utility of reinterpretation-based updating of implicit impressions. In Study 10, I present work testing whether reinterpretation presented in the form of an argument made during a podcast episode on a controversial program for funding conservation efforts to preserve endangered species in Africa – specifically, auctioning off the rights for big-game hunters to legally stalk and kill individual members of the species – is effective in updating implicit impressions of an associated person who is initially viewed in a negative light (a big-game hunter). Additionally, the study measured self-reported reinterpretation to bolster the argument that although the results revealed (expectedly) weaker implicit revision than that observed in earlier studies, reinterpretation played a mediating role in the change that was observed. This study thus begins to build the argument that reinterpretation may represent a more broadly applicable strategy for achieving revision of implicit evaluations toward real targets.

Study 11 examines the generalizability of implicit revision through reinterpretation in another way, by examining whether reinterpretation is still successful when visual features of the target person (which are often thought to be more preferentially impactful on implicit responses; McConnell et al., 2008; Rule, Tskhay, Freeman, & Ambady, 2014) identify the person as a

member of a group (African Americans) stereotyped in a way consistent with the negative (to-be-discredited) first impression. The results find revision toward the target person to be robust, but that updating does not generalize to other members of the same group, thereby showing evidence for a novel form of indirect implicit bias.

Finally, Study 12 closes the chapter by observing how implicit impressions respond to milder, more mundane revelations that earlier expectations about a person were wrong. Specifically, the study takes up a finding reported by Cao and Banaji (2016) that counter-stereotypic information about the professions of two individuals (that a woman named Elizabeth is a doctor and a man named Jonathan is a nurse) only reduced, but did not reverse, implicit gender-based stereotyping of the careers of these individuals. Study 12 adds complexity to that finding by demonstrating an important difference between measures in the extent of implicit stereotyping after such minimal individuating information, and begins to integrate the work with the other studies in this package by finding evidence that factors that increase the depth of processing and psychological realism of the new information – immersion and self-relevance – may hold the key to increasing the efficacy of such minimal, commonplace expectancy violations for enacting reversals of implicit impressions.

#### ***Chapter IV***

In the final body chapter of this work, I present a series of supplementary analyses and additional studies to investigate the implicit nature of the revision effects reported herein, other conditions of automaticity (such as efficiency) under which these effects can emerge, and the robustness of the revision effects to task alterations meant to motivate and facilitate greater participant control over their responses.

The chapter begins with a discussion of prior work on the implicitness of the primary implicit measure used (in some form) in most of the studies in this package. It then presents evidence for a previously unknown feature of the distributions of data produced by the measures in many of the studies in this package that re-raises the question of whether (most of) the responses on this measure are truly implicit, which is followed by new analyses of many of the preceding studies that are intended to find evidence for whether the more unusual data points in the prior studies were generated through explicit responding rather than implicit mechanisms. I also test whether effects previously reported in this package are robust to the exclusion of such data points and whether the unusual data can be accommodated by process models of this measure in a manner that does not discard the assumption of implicitness.

After the new analyses, Chapter IV turns to a set of additional experiments meant to further explore these questions. In Studies 13-14, two different implicit revision effects from earlier studies (one from Chapter II and one from Chapter III) are replicated under conditions that make it easy and encouraged for participants to avoid giving explicit judgments during the implicit task. The findings additionally support the conclusion that participants are not even aware, in the moment, of the critical responses that the task is measuring. In Study 15, a highly distracting and taxing secondary task is added to the implicit measure that partially succeeds in eliminating the distributional anomalies and provides another examination of the efficiency of implicit revision under cognitive load (see also Study 3). Finally, Studies 16-17 will show that when participants are encouraged more blatantly to avoid explicit influence on the implicit measure and given more visually variable stimuli to judge during the task, the distributional anomalies are eliminated, and the critical revision effect is replicated.<sup>1</sup>

---

<sup>1</sup> All studies were supported by a National Science Foundation Graduation Research Fellowship.

## Chapter II. Reversal of Novel Implicit First Impressions<sup>2</sup>

### Abstract

Little work has examined whether implicit evaluations can be effectively “undone” after learning new revelations. Across 7 experiments, participants fully reversed their implicit evaluation of a novel target person after reinterpreting earlier information. Revision occurred across multiple implicit evaluation measures (Experiments 1a and 1b), and only when the new information prompted a reinterpretation of prior learning versus did not (Experiment 2). The updating required active consideration of the information, as it emerged only with at least moderate cognitive resources (Experiment 3). Self-reported reinterpretation predicted (Experiment 4) and mediated (Experiment 5) revised implicit evaluations beyond the separate influence of how thoughtfully participants considered the new information in general. Finally, the revised evaluations were durable three days later (Experiment 6). We discuss how these results inform existing theoretical models, and consider implications for future research.

Key Words: implicit evaluations; attitudes; reinterpretation; AMP; IAT; revision; relevance

---

<sup>2</sup> Published as: Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology*, 108(6), 823–849. <http://doi.org/10.1037/pspa0000021> © American Psychological Association, 2015. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. Please do not copy or cite without author's permission. The final article is available, upon publication, at: <http://doi.org/10.1037/pspa0000021>

What happens when our initial information about someone turns out to be wrong? Can we change how we feel about someone who upon further examination is nothing like our first impressions? Consider the case of a member of the Nazi party in the late 1930's who took over a formerly-Jewish owned business to produce supplies for the German war effort, employing Jews as a cheap source of labor. Learning such details about this person would lead most people to detest him for enabling the Germans during the Holocaust. In the end, however, it turned out that this man – Oskar Schindler – used all his money and connections to keep his Jewish workers from being killed, deliberately producing essentially zero materials for the war effort in his factory and ending up destitute from his efforts to protect his workers (Crowe, 2004; Steinhouse, 1994). As such, Schindler represents a dramatic case of an everyday idea: that people sometimes act with ulterior motives which, when revealed, prompt us to reinterpret their earlier actions. In terms of one of the most basic ways we assess the world around us – evaluation – how do revelations such as these influence us? Are we able to “undo” our first impressions and change our minds about the goodness or badness of others?

The ease of this kind of change may depend on which kind of evaluation is meant: explicit or implicit. *Explicit evaluations* are those measured directly, such as when someone endorses a statement about preference (e.g., “I like her.”). Explicit evaluations seem to be quite capable of reflecting newly learned truths that override earlier information: one can simply choose to reject an old evaluation and endorse a new one (Gawronski & Bodenhausen, 2006). *Implicit evaluations* are those measured indirectly, which means instead of asking people how they feel about a stimulus, the researcher *infers* it by assessing whether the perception of that

stimulus facilitates responses to a different, unrelated stimulus.<sup>3</sup> As it turns out, implicit evaluations are not as easily reversed when prior impressions are found to be false (e.g., Boucher & Rydell, 2012; Gregg, Seibt, & Banaji, 2006; Peters & Gawronski, 2011; see also Wilson, Lindsey, & Schooler, 2000). This work suggests that our implicit first impressions may be relatively hard to undo, persisting even after we learn new information that should override them (see also Rule, Tskhay, Freeman, & Ambady, 2014).

In the present work, we offer a fresh look at this question of whether and how implicit evaluations can be updated to reflect newly learned truths. Given that implicit evaluations uniquely shape and predict behavior (Cameron, Brown-Iannuzzi, & Payne, 2012; Ferguson, 2007; Greenwald, Banaji, & Nosek, in press; Greenwald, Poehlman, Uhlmann, & Banaji, 2009; McNulty, Olson, Meltzer, & Shaffer, 2013; Perugini, Richetin, & Zogmaister, 2010; Towles-Schwen & Fazio, 2006; cf. Oswald, Mitchell, Blanton, Jaccard, & Tetlock, 2013), it is important to know whether – or when – they can be reconciled with one’s reasoning about what is true of the world. In what follows, we describe recent findings that speak to this question, and review what contemporary models of evaluations would predict about this possibility. We then discuss a heretofore little-examined possibility: that, as in the case of Schindler, when new information forces a reinterpretation of the prior impression, reversal of implicit evaluations may be possible.

### **Can implicit evaluations be undone?**

Implicit evaluations were initially assumed to be difficult to change, let alone completely “undo.” They were thought of as the products of long-term exposure to information in one’s

---

<sup>3</sup> Throughout this paper, we use the term *implicit evaluation* because it refers to effects – that is, indirectly-measured unintentional evaluative responses. The term *implicit attitude*, while often used in this literature, is ambiguous in that it might refer either to behavioral effects or the mental constructs posited to explain them (see discussion in De Houwer, Gawronski, & Barnes-Holmes, 2013).

environment (Greenwald & Banaji, 1995) and were assumed to persist in memory even after new attitudes formed (Wilson et al., 2000). More recent work, however, suggests that implicit evaluations *can* sometimes be altered. For example, some have argued that implicit evaluations are enabled by associative processes that entail the spreading of activation through networks based on spatial or temporal proximity, or semantic similarity (Conrey & Smith, 2007; Gawronski & Bodenhausen, 2006; Rydell & McConnell, 2006). Changing these associations has sometimes been assumed to occur through the repeated pairing of the attitude object with counter-attitudinal information (Rydell & McConnell, 2006; Rydell, McConnell, Mackie, & Strain, 2006; Rydell, McConnell, Strain, Claypool, & Hugenberg, 2007).

In line with these assumptions, some researchers have used extensive evaluative conditioning paradigms to try to modify existing implicit evaluations. For example, Karpinski and Hilton (2001) exposed participants to 200 trials of counter-conditioning to try to modify implicit evaluations of the elderly. In this method, change is assumed to depend on the repeated spatial and temporal co-occurrence of the stimulus and evaluative cues (cf. Mitchell, De Houwer, & Lovibond, 2009). If a stimulus, such as a social group, was originally implicitly evaluated as negative, for example, perhaps repeatedly pairing group members with positive cues – without any context, explanation, or reasoning – might nudge the evaluation toward positivity. To be sure, this method has shown that such change of implicit evaluations through evaluative conditioning is possible (e.g., Karpinski & Hilton, 2001; Lai et al., in press; Olson & Fazio, 2006; Rydell et al., 2006).

The question of the current work, however, is whether we can effectively undo implicit evaluations when reasons to doubt our initial impressions come to light. Sometimes we learn new things about the world that immediately transform the meaning of prior knowledge, and to

what degree can such revelations alter our impressions? This kind of learning differs from the mere repeated *pairing* of the attitude object with new information, in that it depends on considerations of truth and falsehood. For instance, new details might emerge about a person suggesting that a first impression of him or her was entirely incorrect, and that some different impression is warranted instead. Beyond adding something new to one's representation of that person, such revelations can suggest that other aspects of an impression should be subtracted. As such, processing the new information might entail figuring out whether to endorse the new information as valid and true, and how the new information is related to older information (Gawronski & Bodenhausen, 2006). Below, we describe theory and empirical work addressing the question of whether implicit evaluations can be “undone” through the affirmation of new impressions as true, the negation of old impressions as false, or the combination of the two.

### **Revision through the addition of new information**

Some studies have tried to change implicit impressions by providing new information about a target that is both different in valence from, as well as totally unrelated to, the initial information (Gawronski et al., 2010; Petty et al., 2006, Study 1; Rydell & Gawronski, 2009; Rydell & McConnell, 2006; Rydell et al., 2007). In these studies, participants are typically first presented with a large number of evaluatively consistent statements about a target person (e.g., “Bob donates his time at the soup kitchen”), and subsequently display the expected implicit and explicit evaluation toward that person. Then, researchers attempt to change overall impressions by presenting new statements with the opposite valence (e.g., “Bob refused to help a child fix his bike”) that are seemingly unrelated to the initial statements. In this task, participants play an active role in affirming the validity of each new piece of information (Rydell & McConnell, 2006), and so the new information is vetted as accurate. But, this approach of adding new



information tends to lead to implicit revision only after considerable amounts of countervailing information is presented, and at a much slower rate than what is needed for explicit revision (e.g., Rydell et al., 2007). To date, the one exception to this is when the new information is extremely negative and rare (Cone & Ferguson, in press). In these studies, after learning a large amount of mildly positive pieces of information about a new person (e.g., “Bob gave a hitchhiker a ride to a shelter”), participants immediately reversed their implicit evaluations of the person after learning a piece of extremely negative and rare information (e.g., “Bob was recently convicted of molesting children”).

Although the evidence for revision through adding new, unrelated information is somewhat mixed, some theories claim that this approach of adding new information is the most likely way in which implicit evaluations can be updated. The Associative – Propositional Evaluation model (APE; Gawronski & Bodenhausen, 2006, 2011) contends that implicit evaluations are generated through associative processing, but can be updated after learning new information that is deemed valid. When people learn about (and believe) new, counter-attitudinal information about a target, for instance, this can create a new counter-attitudinal association, which might then drive the implicit evaluation. Importantly, this model assumes that adding new information in this way is the most likely route to changing implicit evaluations because it is very difficult to overturn, or silence, older associations that were the basis for the initial impression. However, it is not yet clear when such new information, even if fully believed and affirmed, will have a relatively small or large impact on implicit evaluations. Extreme negativity and diagnosticity (Cone & Ferguson, in press) may be one criterion that is necessary for new, unrelated information to lead to revision.

These studies that simply add new, unrelated information to the totality of information about the person (or stimulus) might be classified as “addition” studies. They represent cases where we learn new information about someone that is countervailing to, but independent of, our former impressions. This approach assumes that change will emerge incrementally from the totality of information about the stimulus. In this way, even though adding an extreme piece of information may occasionally swamp out older information (Cone & Ferguson, in press), new information will tend to be added to older information, often resulting in evaluatively complicated representations (i.e., implicit ambivalence; Petty et al., 2006) or contextualized evaluations (Gawronski, Rydell, Vervliet, & De Houwer, 2010; Rydell & Gawronski, 2009), which can allow for the recovery of the initial association with a shift in context (Gawronski et al., 2010; see Bouton, 2004).

### **Revision through the undoing of initial information**

What about when new information forces a change in meaning of the initial information? Is it possible to revise implicit impressions by *undoing* the meaning of previously learned information? In other words, can we effectively “erase” the influence of our implicit associations on the basis of new information? Although some theory maintains that this will be very hard to do (Gawronski & Bodenhausen, 2006, 2011; see also Deutsch, Gawronski, & Strack, 2006; Gawronski, Deutsch, Mbirkou, Seibt, & Strack, 2008), other perspectives claim it is possible. For example, some theoretical work maintains that implicit evaluations are enabled by propositional representations acquired through top-down learning (De Houwer, 2014; Hughes, Barnes-Holmes, & De Houwer, 2011), or are generated through a variety of processes (see, e.g., Amodio, 2014; Amodio & Devine, 2006; Amodio & Ratner, 2011; Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Payne, 2001; Sherman et al., 2008), each of which may

have different capacities for updating (see Amodio & Ratner, 2011). These views either would strongly predict such undoing of implicit evaluations (De Houwer, 2014; Hughes, Barnes-Holmes, & De Houwer, 2011) or are at least open to the possibility (Amodio & Ratner, 2011; Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005).

Another example is the Metacognitive Model (MCM) by Petty and colleagues (Petty & Briñol, 2010; Petty, Briñol, & DeMarree, 2007). In this model, new, countervailing information that overturns older information can indeed change implicit evaluations, but only when the new information is sufficiently elaborated on to *replace* previous associations. Specifically, previously held evaluations that one chooses to reject do not get immediately removed from memory, but instead get tagged as false. These validity tags are initially only weakly associated, which explains why it is difficult to instantly revise (see Petty et al., 2006). However, the model predicts that various factors may moderate the strength of these new “false” tags, such as the extent of elaboration, and the tags may eventually become sufficiently strong so as to prevent the activation of the original evaluation (Petty et al., 2006; Petty & Briñol, 2010).

Despite the theoretical support for implicit revision through the undoing of initial learning, the empirical evidence is mixed. Some studies have attempted to change implicit evaluations by presenting new information about the validity of older information. In such studies, participants first form an initial implicit (as well as explicit) evaluation toward a novel person or group, and then are told that the initial information was true or false (Boucher & Rydell, 2012; Peters & Gawronski, 2011). Asking people to simply “negate” or undo a prior impression in this way has typically been less effective in shifting implicit than explicit evaluations (Deutsch et al., 2006; Gawronski & Bodenhausen, 2011, p. 88; Gawronski et al., 2008; Gregg et al., 2006). The only way that such negation instructions lead to implicit revision

is when they are presented *simultaneously* with the initial information (Peters & Gawronski, 2011), and are sufficiently salient to elicit considerable attention (Boucher & Rydell, 2012).

These studies might be classified as “subtraction” studies because the new information requires people to “unlearn” the prior information, in effect subtracting its influence from their impressions. These might represent cases where we learn that our first impressions were actually based on false rumors, for instance. This kind of subtraction method of telling people to simply reject initial information as false may present several challenges though, which may explain why empirical attempts have yielded only mixed evidence. First, asking people to negate their initial impression may prompt them to try suppressing it, which might ironically keep the rejected thought active (Wegner, 1994). Second, people may be unable to erase all traces of implicit evaluations when instructed to do so, in line with memory work suggesting that intentional forgetting does not erase all traces of a memory (Bjork & Bjork, 2003). Even if participants could, in theory, respond to an instruction to negate everything they have previously learned by thoroughly reinterpreting those details, they may sometimes have low motivation or ability to do so, resulting in insufficiently deep processing of the negation (Boucher & Rydell, 2012; see also Petty et al., 2006). And, finally, learning that initial information was false does not necessarily imply the opposite impression. Learning that someone *did not* perform a negative behavior does not mean the person enacted a positive one, and vice versa. Attempting to silence initial information by classifying it as false would seem to face multiple kinds of challenges, and these challenges might explain the mixed empirical record to date.

### **Joining forces: Revision through subtraction *and* addition**

What happens when we learn new, counter-attitudinal information that *also* overrides the initial impression? That is, what about trying to encompass both an addition as well as a

subtraction approach to implicit revision? For example, learning that Schindler was actually a hero both adds new information as well as *changes* the meaning of the initial information. The fact that he employed Jewish workers in his factory, for example, now has a completely different meaning (and evaluative connotation). Learning new information that also forces a change in the meaning of older information would seem to possess the advantages of addition and subtraction approaches, while avoiding the pitfalls of using either by itself.

To our knowledge, there are only a few studies that provide a test along these lines. Gregg et al. (2006, Studies 3 and 4) attempted this approach in two of their studies. They informed participants that their impressions of two novel groups should be flipped: in one (Study 3), the experimenter had ostensibly made a mistake in informing them of which group was positive and which negative; in the other (Study 4), the groups were said to have changed in their moral character over time, the formerly negative one becoming positive, and the formerly positive one becoming negative. These seem like “subtraction + addition” methods in that the new information states that the evaluation attached to each of the groups should be reversed (e.g., the “Niffites” are no longer bad, and are instead now good), but, in this case, they did not find any evidence of implicit revision. However, these particular instantiations of a “subtraction + addition” approach may not have been ideal. First, in Study 3, the new information did not change the *meaning* of the groups’ initial behaviors so much as switch the authorship of those behaviors. The behaviors had the same evaluative connotation, but the mapping of behavior to group was supposed to be switched. Secondly, whether people are able to revise their impressions would seem to depend critically on the believability of such a switch. If people find the notion of a group completely switching its entire moral character unlikely (which is what Study 4 asked participants to believe occurred over time), then we might not see implicit revision

even if they had been able to do so (especially given the stickiness of initial immorality; e.g., Knobe, 2006; Malle & Knobe, 1997; Reeder, Pryor, & Wojciszke, 1992). Finally, these concerns might have been especially pronounced because the participants in both studies knew that the groups were fictional and the scenarios hypothetical. In Study 3 for instance, they learn that for some other participants the goodness or badness of the groups is the reverse of what they were told, which might privately undermine their sense that either group is “truly” good or bad. Explicit evaluations, on the other hand, may have reflected participants’ perceptions of the expectations of the experimenter.<sup>4</sup>

A better test of the “subtraction + addition” approach might be to present new countervailing information that truly changes the meaning of the old information, and to do so in a way that has some ecological validity (i.e., use a paradigm that maximizes believability). To our knowledge, there is only one study that has tested such an approach. Wyer (2010; Study 2) showed that implicit evaluations of a novel target were revised in light of new, counter-attitudinal information that *changed* the interpretation of prior details – an apparent skinhead who behaved in an off-putting way turned out to be ill with cancer. This changed participants’ implicit evaluation of the target, but only if they were able to revisit each one of his initial behaviors once they get the new information. Wyer suggested that this focused rehearsal may have sufficiently strengthened the “false” tags linked to those prior details to allow the implicit revision, in line with the MCM perspective (Petty et al., 2006). Although there remain many

---

<sup>4</sup> Note that a few studies similar to those reported by Gregg et al. (2006) did not technically measure implicit evaluations. Wilson and colleagues (2000) employed a measure requiring rapid explicit judgments, and compared these to more conventional slower explicit judgments. Petty et al. (2006) employed a study design (in Experiment 2) asking participants to switch the targets of two sets of information, but the implicit measured tapped associations between the targets and confidence vs. doubt, rather than evaluations. As such, very few studies have attempted “addition + subtraction” designs while measuring changes in implicit evaluations.

questions about why and how this effect emerged, it is intriguing, and raises the possibility that when people learn information that makes them *reassess* their prior knowledge, their implicit evaluations can be updated accordingly. In what follows, we consider this possibility.

### **Reinterpretation**

Although existing theories are open to the possibility of undoing implicit evaluations, in line with some supportive findings, the mechanisms through which it might occur remain largely unknown. We propose that the ability of new information to *recast* the old information on which the initial evaluation was based – such as in the case of Oskar Schindler – is one mechanism of change that may be especially effective. In particular, when new information prompts a reinterpretation (i.e., a change in the evaluative meaning) of old information, we predict that implicit evaluations will be updated accordingly.

Reinterpretation of prior information may be uniquely positioned to produce strong revision of implicit evaluations. This strategy involves not just the invalidation of the initial impression (i.e., subtraction), but also the *replacement* of that impression with a countervailing other (i.e., addition), and it does this in one fell swoop. That is, it introduces an explanation for why previous learning should be reinterpreted, and revised in the opposite evaluative direction. This may often be more effective in producing revision of implicit evaluations than either addition or subtraction approaches alone. If there are reasons to suspect that both rejecting a prior impression and affirming a new one may have limitations when implemented separately (as we previously reviewed), the initial demonstration in Wyer (2010) is a promising sign that reinterpretation – a change in the meaning of earlier details such that an initial impression is both negated *and* replaced with another – may be an effective way to overturn implicit evaluations.

Though the demonstration in Wyer (2010) remains the best test in the literature of the possibility that reinterpretation may produce revision of implicit evaluations, there is much that remains unknown. Most critically, because there were no measures of the nature of the thinking done by participants or comparisons of the effectiveness of different types of counter-attitudinal information, it is unclear whether there was any reinterpretation at all; participants given the new information may have simply elaborated on it as they revisited the initial information without changing their understanding of the initial information. In other words, there was no evidence about the process leading to the revision, and whether it involved any reassessment of the meaning of the initial behaviors. It may be that elaborating on *any* counter-attitudinal new information, without a rejection of the earlier impression, would have been sufficient.

In the current work, we examine whether implicit evaluations can be fully and durably reversed when new information *changes the meaning of a previous impression*. This would demonstrate that implicit evaluations can be fully reversed after reasoning about a prior impression. In addition, we identify the process by which this kind of change occurs, and reveal its operating characteristics: it requires more than the simple pairing of the old attitude object with counter-attitudinal information, and is not reducible to more extensive *general* thinking about that information; rather, it occurs through reinterpretation specifically.

### **Overview of Current Work**

We developed a new paradigm tailored to this research topic. In each experiment, participants read a story, presented one sentence at a time, about an individual named Francis West who is described as breaking into and causing damage to two homes. Participants' implicit evaluations toward Francis West are then measured. Afterwards, they read a final piece of information which either maintains the gist of what they already read (control conditions in



which Francis remains negative) or dramatically reverses it (experimental condition in which Francis becomes positive) by offering a reinterpretation of what was previously learned: The houses were on fire, and Francis was searching for two young children who he knew were inside.

With this paradigm, we accomplish a number of goals across the studies. First, we use a paradigm in which participants are learning about an ostensibly real person through a cohesive narrative that is meant to be engaging and credible. In this way, we hope to maximize participants' motivation and attention while in the learning paradigm. Secondly, we demonstrate fast revision in the direction of negative to positive, which has been shown to be especially difficult in recent work (Cone & Ferguson, in press) possibly due to negativity dominance (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Cacioppo, Gardner, & Berntson, 1997; Rozin & Royzman, 2001). Thirdly, the paradigm leads participants to form implicit evaluations toward a novel person, which enables us to test the ways in which implicit evaluations can be *changed* (Fazio, 2007; Ferguson & Fukukura, 2012; Gregg et al., 2006). That is, we can ensure that we measure learning, rather than re-activation of previously learned material (e.g., see Gregg et al., 2006, pp. 16). Presenting counter-attitudinal information about familiar objects might simply activate prior learning, the characteristics of which (how long it took to learn it, etc.) would be unknowable.<sup>5</sup>

In Experiments 1a and 1b, we demonstrate that reversal occurs in this paradigm using two different implicit measures of evaluations. Experiment 2 shows that the relevance of the

---

<sup>5</sup> Presenting participants with counter-attitudinal information about a known social group might indeed result in changed implicit evaluations of the group. But, this does not mean that participants *learned something new*. Thus, the evidence for the context-dependence of implicit evaluations (e.g., Blair, Ma, & Lenton, 2001; Dasgupta & Greenwald, 2001; Ferguson & Wojnowicz, 2011; Wittenbrink, Judd, & Park, 2001; see Blair, 2002, for a review) does not address the topic of how easily implicit evaluations can reflect new information.

new information in recasting the old is essential to this reversal by comparing it with a condition containing equally positive but irrelevant information. Experiment 3 examines whether the revision occurs through an active thought process of reappraising the old information, by manipulating cognitive load to test whether a reduction in cognitive resources would undermine revision. Experiment 4 demonstrates that participants' self-reported belief that the new information changes the meaning of the prior story predicts the degree of revised implicit evaluations of Francis West, even when controlling for more general measures of the speed of thinking and extensiveness of thinking. In Experiment 5 we show that self-reported reinterpretation of the prior story mediates the effect of the new information on implicit evaluations, even when controlling for the extent to which participants reported thinking about the new information in general. Finally, in Experiment 6 we demonstrate the durability of the revised implicit evaluations over three days. For all studies, we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study (Simmons, Nelson, & Simonsohn, 2012).

### **Experiment 1a: From Burglar to Hero – I**

We presented participants with a story in which a novel person ("Francis West") is depicted breaking into and causing great damage to two houses on his street, followed by a revelation (in the experimental condition) that his behavior was motivated out of a desire to save children inside from a fire spreading through the houses. Thus, this revelation implies a reinterpretation of the prior story. In a control condition, participants instead read one additional piece of information that does not contradict, but rather is consistent with, the information learned before (see Appendix). In addition to measuring implicit evaluations, we assessed

explicit evaluations and a variety of participants' reactions to the story, including story comprehension, confusion, and how deeply they reported thinking about the story's details.

## **Method**

**Participants.** Two hundred workers on Amazon's Mechanical Turk website (www.mturk.com) participated in the study for \$0.75 (55% male;  $M_{\text{age}} = 37$  years,  $SD = 14$ ). We selected this number a priori so as to collect approximately 100 per between-participants condition. Because this was our first attempt at testing the effect, we collected enough data to be able to detect a moderately sized effect.

**Materials.** In order to induce an initially negative implicit evaluation toward a novel target person, participants were led through a story detailing supposedly true events centering around an individual named Francis West. The story was presented in a linear piecemeal fashion, across 26 screens. The described events portray Francis West as a neighbor who ransacks the homes of his neighbors, destroying their property (e.g., throwing a pot of water onto a laptop) and taking "precious things" from the bedrooms. After initial assessment of implicit evaluations toward Francis, participants were then presented with a single screen of additional information that varied by condition. In the control condition, the new information continued the thread of the story: Francis began to chuck rocks at the windows of the houses he had just pillaged. In the other condition (henceforth dubbed the "fire rescue" condition), participants instead read that Francis broke into the houses because he saw that they were on fire, and the only precious things he removed from the bedrooms were the young kids of the families.

As participants read each statement about Francis in both the initial and subsequent story periods, an image of the upper body of a white male labeled "Francis West" was displayed on the screen. Each participant was randomly assigned an image of one individual to serve as Francis

West, out of a set of 11 such images drawn from previous research (Minear & Park, 2004). The men in all photographs had neutral expressions, and ranged in age from 20 to 33 years.

***Implicit Evaluations.*** Implicit evaluations were measured twice, once after reading the initial statements (Time 1) and once after reading the final information (Time 2). The Affect Misattribution Procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005) was administered at both instances (see the *Procedure* section). Each AMP consisted of 40 trials, with separate sets of Chinese characters used at Time 1 and Time 2. The order in which the sets were administered was counterbalanced across participants.<sup>6</sup> The image of Francis West that was assigned to the participant served as the prime on half of the trials, and each of the other 10 images of unknown individuals served as a prime on two of the other trials. Participants were instructed that they would sequentially view a set of Chinese ideographs, and that their task was to determine for each whether it was more or less pleasant than the average ideograph by pressing the *k* and *d* keys on their keyboard, respectively. On each of the 40 trials, participants were first presented with a prime photograph of Francis or an unknown individual for 75 ms, followed by a blank screen for 125 ms, an ideograph for 100 ms, and finally a pattern mask of black and white noise, which remained on the screen until the participant responded. Participants were told that though the images that precede the ideographs may sometimes be positive or negative, they were to prevent these images from affecting their ratings, and instead should evaluate the ideographs solely on their own merits. Previous research suggests that this measure taps evaluative reactions toward the primes that are misattributed to the relatively neutral targets, thus providing

---

<sup>6</sup> The order in which the two sets of ideographs were administered will not be discussed further; order affected only one analysis of interest across all six studies. In Experiment 3, the interaction of time, prime person, and story condition was significantly moderated by ideograph order, such that the revision effect was stronger in one counterbalance condition than in the other (but the revision effect, and all simple effects of interest, were still significant in both). Ideograph set order had no similar effects in any other analysis.

a measure of spontaneously elicited, unintentional evaluations of the primes (Payne et al., 2005; Payne et al., 2013; cf. Bar-Anan & Nosek, 2012).

**Explicit Evaluations.** Explicit evaluations toward Francis were measured at Time 2 using measures adapted from previous research (Rydell & McConnell, 2006; Rydell, McConnell, Mackie, & Strain, 2006). Participants indicated how likable Francis is from 1 (*very unlikable*) to 7 (*very likable*), and completed 7-point semantic differential scales in random order on the dimensions of bad-good, mean-pleasant, disagreeable-agreeable, uncaring-caring, and cruel-kind. These six items were reliable,  $\alpha = .994$ , and were combined into a single scale for the analyses.

**Questionnaire Measures.** Participants completed three multiple choice questions asking questions about the story that have different probable answers in the control and fire rescue conditions: why Francis threw water around the house (e.g. to ruin items, to put out a fire), why the cat died (e.g. Francis stepped on it, smoke inhalation), and what he removed from the houses (e.g. jewelry, children). They then identified Francis in a lineup of the 11 faces presented in the study (1 Francis and 10 control), indicated their level of confusion about what happened in the story on a scale from 1 (*Not confused at all*) to 7 (*Completely confused*), the extent to which they thought about the Time 1 story elements after reading the new information at Time 2 on a scale from 1 (*Not at all*) to 7 (*A lot*), and how hard it was to make sense of how the Time 2 information fit with the rest of the story on a scale from 1 (*Not at all hard*) to 7 (*Very hard*).

**Procedure.** After providing informed consent, participants were instructed to read each statement in the story depicting the initial events surrounding Francis West. They were told to pay careful attention to the details that unfolded, as they would be asked questions about their perspective on the events later in the study. Participants proceeded through the screens containing the descriptions at their own pace, but spent a minimum of 3 seconds on each screen.

After reading the initial story about Francis West, participants took the first AMP. Then, they were informed that they would read a final piece of information about the events described previously, and were shown either the control or fire rescue information. They were instructed to think about how this information relates to what they learned before, and were required to wait at least 15 seconds before advancing. They then completed the Time 2 AMP. Next, they indicated whether they knew Chinese, completed the explicit evaluation scale, and answered the other questionnaire items. Finally, they completed demographic questions, a set of measures unrelated to the present study regarding political evaluations, and were debriefed and compensated.

## Results

**Data preparation.** Following Payne et al. (2005), data from 4 participants were excluded for indicating familiarity with Mandarin or Cantonese (2% of cases) and 18 participants for using a single key on every trial of at least one AMP, indicating a disregard for the instructions (9% of cases). This left 178 cases for the final analysis.

**Comprehension checks.** Participants in both conditions showed good comprehension of the story details, with 93% in the control condition and 84% in the fire rescue condition answering the three interpretative questions in a manner fully consistent with the condition to which they had been assigned. This difference was marginally significant,  $\chi^2 = 2.94, p = .086$ . In addition, every participant correctly identified Francis West in the lineup of 11 photographs.<sup>7</sup>

---

<sup>7</sup> For descriptive purposes, we continued to collect information on how frequently participants completed the inferential items in a manner fully consistent with their story condition, and whether they could correctly identify the face of Francis West out of a lineup, throughout all of the subsequent studies. From Experiment 1b onward, we also always included a simple manipulation check item asking participants to identify from a short list which final information had been presented to them about Francis West. We never made a priori predictions about these measures, and so do not discuss them further in this paper, though the correlations between implicit liking and the inferential items in the fire rescue condition are available in Table 1. We chose in advance to include participants regardless of their performance on these measures.

**Implicit evaluations toward Francis West.** Implicit evaluations were assessed using a 2 (Measurement Time: time 1 and time 2) x 2 (Prime Person: Francis West and control faces) x 2 (Story Condition: control or fire rescue) mixed design, with the first two factors manipulated within-participants and the third manipulated between-participants. The proportion of trials in each cell of this design on which participants indicated that the Chinese ideograph was more pleasant than average served as the dependent variable in a repeated-measures ANOVA.

Every effect in the model was significant, including our predicted 3-way interaction between measurement time, prime person, and story condition,  $F(1,176) = 36.551, p < .001, \eta_p^2 = .172$ . Decomposition of this interaction revealed that in the control story condition, in which Francis is depicted as negative at both time 1 and time 2, no interaction between time and prime person emerged,  $F(1,176) = 1.55, p = .214$ . Instead, as predicted, there was solely a main effect of prime, such that implicit positivity toward Francis was lower ( $M = .40, SD = .27$ ) than implicit positivity toward the neutral faces ( $M = .63, SD = .19$ ),  $F(1,176) = 50.51, p < .001, \eta_p^2 = .223$ . In the fire rescue condition, however, there was the predicted significant interaction between measurement time and prime person,  $F(1,176) = 51.90, p < .001, \eta_p^2 = .228$ . At time 1, Francis was less positive ( $M = .40, SD = .30$ ) than the neutral faces ( $M = .60, SD = .19$ ),  $F(1,176) = 23.53, p < .001, \eta_p^2 = .118$ , whereas at time 2, Francis was more positive ( $M = .72, SD = .22$ ) than the neutral faces ( $M = .54, SD = .23$ ),  $F(1,176) = 17.42, p < .001, \eta_p^2 = .09$ . Viewed another way, implicit positivity toward Francis did not show a shift from time 1 to time 2 in the control story condition,  $F(1,176) = 1.43, p = .234$ , but showed a significant increase from time 1 to time 2 in the fire rescue condition,  $F(1,176) = 84.58, p < .001, \eta_p^2 = .325$  (see Figure 1).

---

Across the samples of all of our experiments, after setting aside excluded participants (reasons reported in text), 98.73% were correct in identifying the final information they had been shown, 99.58% correctly identified the image of Francis West, and 82.70% responded to the inferential questions in a manner fully consistent with their story condition.

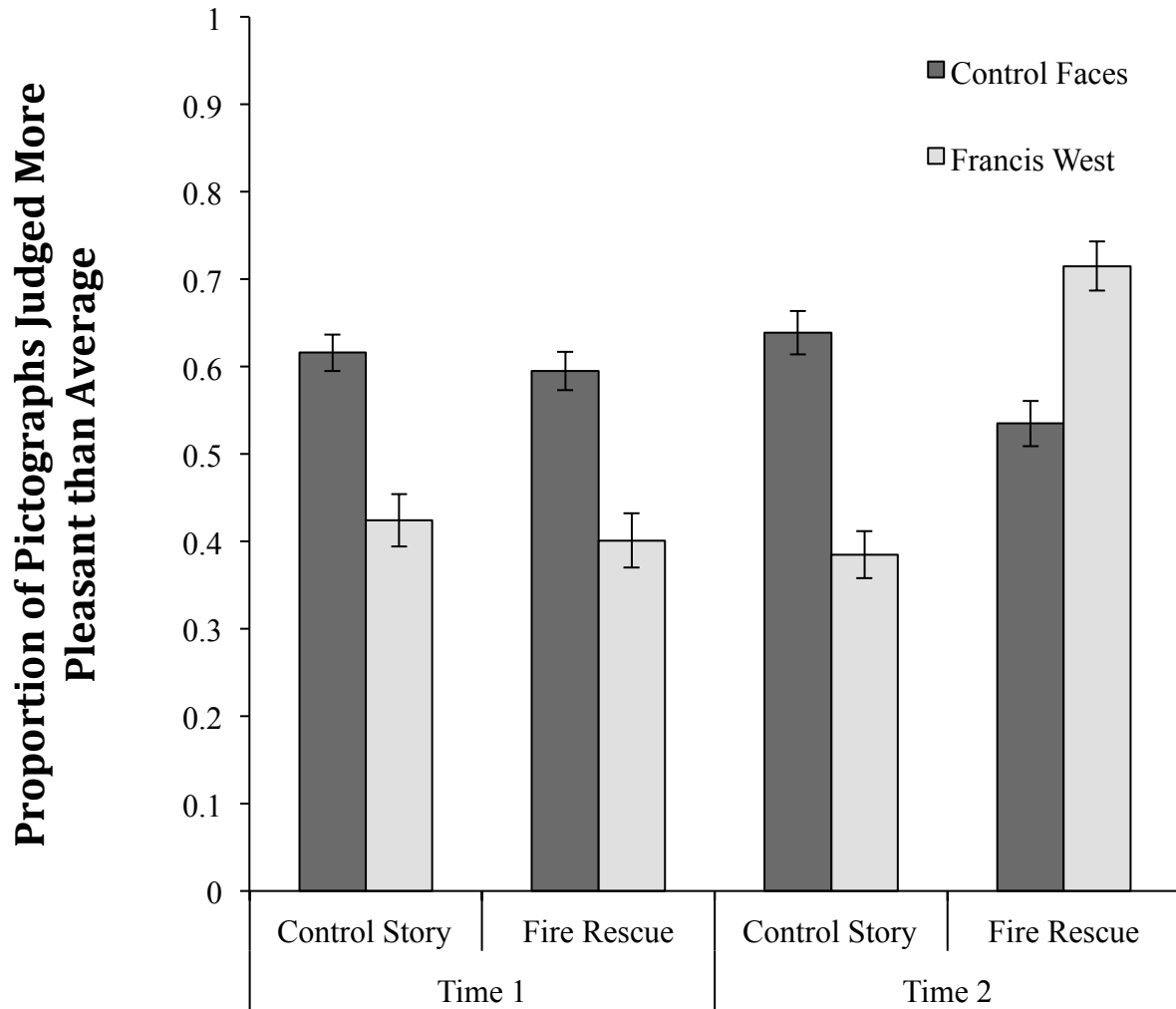


Figure 1. Mean proportion of ideographs judged more pleasant than average in Experiment 1a, by face prime, time, and story condition. Error bars are standard errors.

**Explicit evaluations toward Francis West.** Explicit evaluations were measured after the second AMP, and differed significantly by story condition, unequal variances  $t(103.36) = 34.77$ ,  $p < .001$ , Cohen's  $d = 6.84$ . Participants had more positive explicit evaluations toward Francis West in the fire rescue ( $M = 5.99$ ,  $SD = 1.21$ ) versus control condition ( $M = 1.20$ ,  $SD = .41$ ).



Finally, rather than reporting the correlations between implicit and explicit evaluations (as well as their relations to other measures) in the main text for each study, we summarize this information in Table 1.

**Story condition effects on other questionnaire measures.** Participants reported significantly more confusion with the story in the control ( $M = 3.64$ ,  $SD = 2.05$ ) versus fire rescue condition ( $M = 1.93$ ,  $SD = 1.57$ ), unequal variances  $t(169.50) = 6.27$ ,  $p < .001$ , Cohen's  $d = .96$ . This pattern was the same for reported difficulty making sense of the final story information, unequal variances  $t(175.96) = 2.95$ ,  $p = .004$ , Cohen's  $d = .44$ . This result makes sense, given the lack of resolution available to participants in the control condition regarding the motivation for the destruction perpetrated by Francis West. Also expected, there was a strong story condition effect on the self-reported extent to which participants thought about the overall story upon reading the final information, with those in the fire rescue condition reporting more thinking ( $M = 6.26$ ,  $SD = 1.20$ ) than those in the control condition ( $M = 5.12$ ,  $SD = 1.72$ ), unequal variances  $t(163.02) = 5.14$ ,  $p < .002$ , Cohen's  $d = .80$ . However, none of the above questionnaire items moderated or mediated implicit evaluation revision, and the three-way interaction between time, target person, and story condition remained even when controlling for any of these (mean centered; Aiken & West, 1991).

Table 1. *Correlations Between Implicit and Explicit Evaluations and Questionnaire Measures in Fire Rescue Condition at Time 2, Experiments 1-6*

	Implicit Evaluations							
	Study 1a	Study 1b	Study 2	Study 3	Study 4	Study 5	Study 6	
							Time 2	Time 3
Perfect comprehension	.12	.10	-.07	.27**	-.03	.08	-.06	.14
Confusion	-.04	-.12	.06	.02	-.03	-.24**	-.03	-.14
Difficulty making sense	.04		-.02					
Extent of thinking	-.03		.18†					
Extent of thinking (new)				.19*				
Extent of thinking (old)				.02				
Subjective meaning change					.31*	.23**		
Rapid vs. gradual thinking					-.10			
Extensiveness of thinking (Studies 4-5)					.01	-.09		
Positive mood							.28**	.15
Belief that story is real							.09	.14
	Explicit Evaluations							
Perfect comprehension	.32**	.48***	.54***	.52***	.40***	.30***	.19†	.22*
Confusion	-.33**	-.58***	-.55***	-.45***	-.41***	-.45***	-.26*	-.23*
Difficulty making sense	-.31**		-.65***					
Extent of thinking	.26*		.18†					
Extent of thinking (new)				.32***				
Extent of thinking (old)				.05				
Subjective meaning change					.71***	.58***		
Rapid vs. gradual thinking					-.40***			
Extensiveness of thinking (Studies 4-5)					-.26*	-.10		
Positive mood							.50***	.42***
Belief that story is real							.19†	.16
<b>Implicit evaluations</b>	.30**	.16†	.14	.32***	.24†	.48***	.27**	.35**

*Note.* Cell values are Pearson correlations. In Experiments 1a and 2-6, the correlations involving implicit evaluations are partial correlations with the proportion of ideographs on Francis trials judged more pleasant than average, controlling for the proportion of ideographs on Control trials

judged more pleasant than average. The exception is the IAT used in Experiment 1b, which is a relative measure of positivity toward Francis (vs. control faces); thus, the *D* scores were used, without any covariates. On all implicit and explicit liking measures, higher scores indicate more positive evaluations toward Francis West.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ . †  $p < .1$ .

## **Discussion**

Experiment 1a shows that participants strongly revised their implicit evaluations toward a novel target person, once given a reinterpretation of the original information. Whereas those in the control condition showed a persistence of their initial, negative evaluations, those in the fire rescue condition switched to significant implicit positivity toward Francis West after reading the revelation. Experiment 2 will begin to examine mechanism, but first Experiment 1b replicates the basic pattern with a different implicit evaluation measure.

### **Study 1b: From Burglar to Hero – II**

In Experiment 1b, we sought to replicate the revision effect using the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998). In addition, in order to induce an even stronger initial negative evaluation (and hint at a possible motive in the control condition), we added a new detail to suggest in the first part of the story that Francis' actions were a hate crime against the town's first interracial families in the neighborhood. We anticipated that the revelation that he actually was saving the children from a fire would effectively counter this suspicion and produce reversal just as it had in Experiment 1a. As a final addition, we measured explicit evaluations after both IATs. This allows us to show how implicit liking of Francis West is indeed changing in a similar way to explicit liking.

## Method

**Participants.** Three hundred and one participants recruited from Mechanical Turk were paid \$1.00 to participate in this study (55.6% male;  $M_{\text{age}} = 32.0$  years,  $SD = 11.8$ ). We sought to collect data from approximately 150 participants for each of the two between-participants story conditions, and so intended to collect data from 300 participants. The number is more than in Experiment 1a because we thought that the IAT, which relies on response times, might produce noisier data than the AMP used in Experiment 1a. One additional participant completed the study without subsequently submitting it for compensation.

**Materials.** The story that participants read at time 1 was identical to that from Experiment 1a, except for the addition of a new event at the very start: “Francis' small town was about 99% white, but recently the Griffins and Wards, the town's first ever interracial families, had moved in.” By providing the hint of a motive for Francis' actions, we thought that this might reduce the persistent difference in expressed confusion between the two conditions, as well as provide even more initial negativity toward Francis that must later be overcome.

**IAT.** The IAT included the same eleven faces used in the AMP in Experiment 1a, with one of the eleven having been randomly assigned to be used as Francis West for each participant. Positive adjectives presented during the IAT included *wonderful*, *excellent*, *good*, *great*, *appealing*, *outstanding*, *lovely*, *fantastic*, *beautiful*, and *amazing*. Negative adjectives included *horrible*, *terrible*, *awful*, *disgusting*, *offensive*, *hideous*, *revolting*, *bad*, *dreadful*, and *nasty*.

During the critical blocks, participants quickly sorted stimuli into one of four categories: *Francis West*, *Other People*, *Good Words*, and *Bad Words*. At any one time, two of these category labels (one person category and one adjective category) were displayed on the left side of the screen, and the other two were presented on the right side. A single stimulus from one of

the four categories appeared on each trial. Half of the person trials consisted of an image of Francis West, while the other half displayed randomly chosen images from the other 10 images (control faces). Likewise, on half of the adjective trials a positive adjective was randomly selected from the list, while on the other half a negative one was. Each stimulus was displayed until the participant registered a response by pressing one of the two keys. If the response was correct, the next trial began; if it was incorrect, a red “X” appeared on the screen until the participant gave the correct response. The intertrial interval was 250 milliseconds. The IAT consisted of seven blocks, with 20 trials in practice blocks 1, 2, 3, and 6, and 40 trials in test blocks 4 and 7 as well as transition practice block 5 (for details, see Greenwald, Nosek, & Banaji, 2003). The order of the Francis West + Bad and Francis West + Good blocks was counterbalanced across participants.

**Procedure.** Participants first completed the story procedure in the same fashion as in Experiment 1a. Then, they took the first IAT and the same explicit evaluation items from Experiment 1a. Participants were next presented with the time 2 story information in the same manner as before, either the fire rescue or control information. Immediately thereafter they completed the second IAT, the explicit evaluation scale, the story comprehension items, the photo identification, the confusion items, a new multiple-choice manipulation check question asking them to directly identify the final story information that they had been presented with, and demographic questions. They were then debriefed, thanked, and compensated for their time.

## **Results**

**Data preparation.** We calculated implicit positivity toward Francis West for both IATs according to the *DI* scoring algorithm (Greenwald, Nosek, & Banaji, 2003). The differences between blocks were computed so that higher scores meant faster responding during the “Francis

West + Good” pairing. On this measure, more positive scores are taken to suggest more implicit positivity toward Francis West, and more negative scores are taken to suggest more implicit negativity toward him, relative to the control faces. Due to server error, IAT data were not recorded for 8 participants. Eleven participants were excluded for responding faster than 300 ms on over 10% of trials, following scoring recommendations (Greenwald et al., 2003), leaving the final sample with 282 cases (54.6% male;  $M_{\text{age}} = 32.2$  years,  $SD = 11.8$ ).

**Implicit evaluations toward Francis West.** We analyzed implicit positivity toward Francis West by submitting  $D$  scores to a 2 (Measurement Time: time 1 and time 2) x 2 (Story Condition: control or fire rescue) mixed ANOVA, with the first factor manipulated within-participants. This analysis revealed a main effect of time, qualified by the time by story condition interaction,  $F(1,280) = 10.65, p = .001, \eta_p^2 = .037$ . Simple effects tests demonstrated that, as expected, implicit evaluations toward Francis did not vary by story condition at time 1,  $F(1,280) = 1.11, p = .292$ , but were significantly negative ( $M = -.066, SD = .394$ ) across the sample, one-sample  $t(281) = -2.82, p = .005$ , Cohen’s  $d = .336$ . At time 2, however, implicit positivity of Francis was higher in the fire rescue condition ( $M = .197, SD = .34$ ) than the control story condition ( $M = -.021, SD = .33$ ),  $F(1,280) = 45.44, p < .001, \eta_p^2 = .140$ . Although implicit evaluations toward Francis were significantly greater than zero in the fire rescue condition at time 2, one sample  $t(147) = 7.04, p < .001, d = 1.16$ , they were no longer significantly below zero in the control condition, one sample  $t(133) = -.72, p = .474$ . Thus, the significant change in positivity across time in the fire rescue condition paralleled the shift observed in Experiment 1a,  $F(1,280) = 45.44, p < .001, \eta_p^2 = .140$ , while a marginal shift in the positive direction also occurred in the control condition that was not present in the prior study,  $F(1,280) = 3.64, p = .057, \eta_p^2 = .013$  (see Figure 2 for the mean  $D$  scores in each condition).

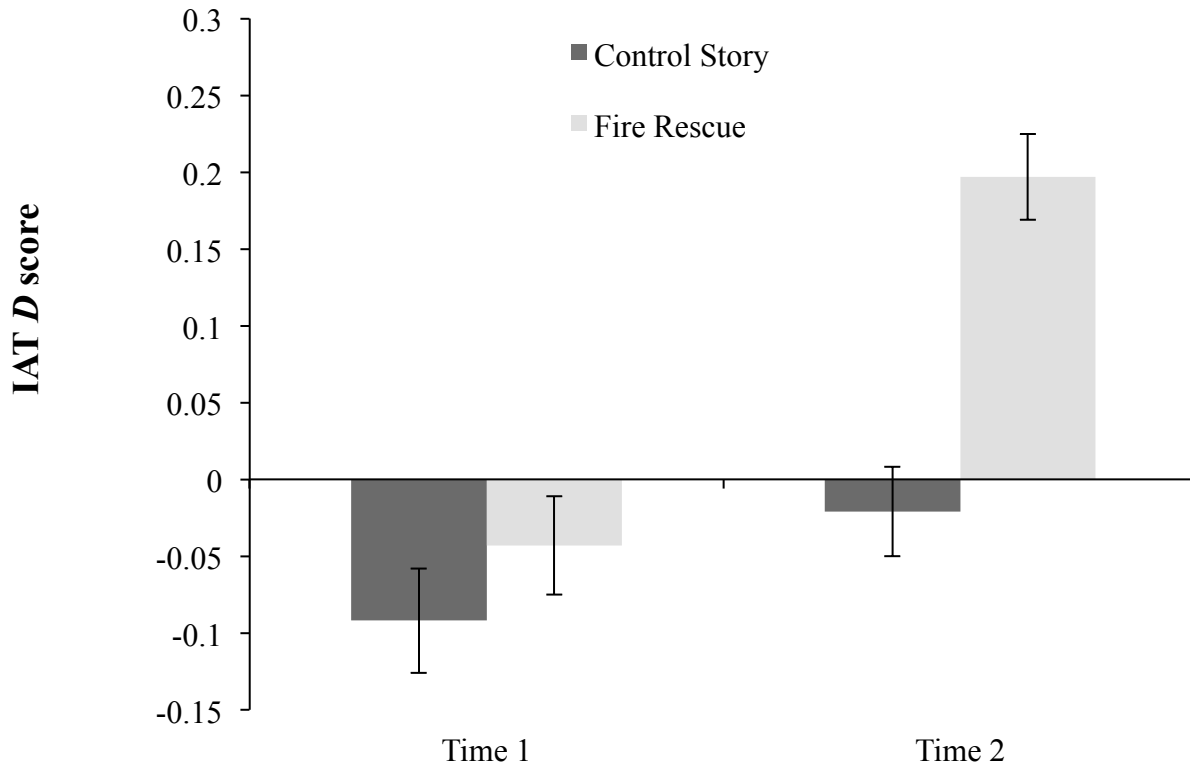


Figure 2. Mean *D* scores in Experiment 1b, by measurement time and story condition. Error bars are standard errors.

**Auxiliary analysis.** The shift toward neutral implicit evaluation of Francis West from time 1 to time 2 in the control condition might suggest that a portion of the sample is relatively unconvinced of Francis West’s badness at time 1 and, after being similarly unconvinced at time 2, lose whatever tenuous negativity toward him they might have possessed at time 1, producing the condition’s overall shift to neutral at time 2. In order to show reason-based revision, we need to demonstrate that for those individuals that *are* induced with an initial evaluation, new information can quickly and significantly reverse it. Otherwise, a finding that the group as a whole exhibits reversal at time 2 could be a product of shifts in these unconvinced participants, rather than shifts in those who acquired the initial evaluation. As such, we undertook a more

conservative auxiliary test of the revision hypothesis by examining the means of time 2 IAT scores solely among those participants who displayed initial negativity at time 1 ( $D$  scores less than zero). In this subsample, IAT scores at time 2 differed between story conditions,  $t(159) = 5.73, p < .001$ , Cohen's  $d = .91$ , such that scores were still significantly below zero in the control story group ( $M = -.14, SD = .31$ ), one sample  $t(76) = -3.80, p < .001$ , and were now significantly above zero in the fire rescue group ( $M = .16, SD = .34$ ), one sample  $t(83) = 4.32, p < .001$ .<sup>8</sup> Although like any measure  $D$  scores are not a process-pure measure of underlying evaluations (Conrey et al., 2005) and thus their absolute values and even rank order are subject to extraneous variation, this time 2 pattern of only those individuals below zero on the IAT distribution at time 1 corroborates our prediction and demonstrates continued deviation from zero in the control story. This even more conservative test of our revision hypothesis builds support for the theory and corroborates the pattern from Experiment 1a (the results reported in Experiment 1a hold as well when submitted to this same analysis).

**Explicit evaluations toward Francis West.** Explicit evaluations toward Francis West were measured at time 1 and time 2. Analysis using a 2 (Measurement Time: time 1 and time 2) x 2 (Story Condition: control or fire rescue) mixed ANOVA revealed that a main effect of measurement time was qualified by the expected time X story condition interaction,  $F(1,280) = 1030.93, p < .001, \eta_p^2 = .786$ . Simple effects tests revealed that explicit liking of Francis West increased from time 1 ( $M = 1.30, SD = .71$ ) to time 2 ( $M = 5.93, SD = 1.47$ ) in the fire rescue

---

<sup>8</sup> This analysis differentially includes participants who completed the compatible-first order (Francis+bad) over the incompatible-first order (Francis+good), as the compatible-first order tends to produce more negative scores. However, we find the same results when splitting by order:  $D$  scores become positive in the fire rescue condition in both orders,  $M = .16, SD = .33, t(43) = 3.21, p = .003$ , and  $M = .16, SD = .35, t(39) = 2.86, p = .007$ , while  $D$  scores are still negative in the control story condition in both orders,  $M = -.12, SD = .31, t(50) = -2.78, p = .008$ , and  $M = -.16, SD = .32, t(25) = -2.60, p = .015$ , respectively.



condition,  $F(1,280) = 2129.35, p < .001, \eta_p^2 = .884$ . No such change from time 1 ( $M = 1.17, SD = .46$ ) to time 2 ( $M = 1.12, SD = .43$ ) took place in the control group,  $F(1,280) = .17, p = .680$ .

**Story condition effects on other questionnaire measures.** As in Experiment 1a, confusion with the story was higher in the control condition ( $M = 2.79, SD = 1.67$ ) than in the fire rescue condition ( $M = 2.19, SD = 1.62$ ),  $t(281) = 3.07, p = .002$ , Cohen's  $d = .365$ . The addition of the hate-crime element to the story did not appear to stem this trend of higher confusion in the control condition. But once more, controlling for (centered) confusion did not reduce the time X condition interaction to nonsignificance,  $p = .007$ .

## Discussion

We replicated reason-based revision using a different implicit measure. Whereas participants' negative implicit evaluations toward Francis West were not qualified by time in the control condition, those in the fire rescue condition showed a significant reversal from negative to positive. Those positive evaluations in the fire rescue condition after receiving the final information were also significantly more positive than the neutral point; however, the final negative evaluations in the control condition did not differ from neutral overall, departing from the results in Experiment 1a with the AMP. A follow-up analysis, however, showed that among those participants most critical for the demonstration of revision – those who *did* show an initial negative implicit evaluation – negative evaluations persisted in the control group and were reversed in the fire rescue group.

As of yet, our results do not yet address the *how* or *why* of implicit evaluation revision. Our argument is that it is the relevance of the fire rescue information to the initial story that prompts revision through reinterpretation. The story presented in the Francis West paradigm appears to fit this bill, but the work to demonstrate the operative mechanisms has yet to be done.

In fact, the failure of self-reported extent of thinking to moderate revision in Experiment 1a could be seen as an initial point against this idea, if reinterpretation of prior story events requires any amount of deliberation about the story, as we suspect that it might. However, this single questionnaire item may not adequately tap the degree to which reinterpretation took place, either because introspective access to this mechanism is weak (Nisbett & Wilson, 1977), or the reinterpretation that is required is easy to execute in this particular case. As such, we examine the conditions under which reason-based implicit evaluation change occurs in Experiment 2.

## **Experiment 2**

To identify whether reinterpretation of the initial information is critical for revision, we compared evaluation change in the fire rescue condition with another condition in which extremely positive information about Francis is presented, but does not prompt a reinterpretation of Francis' initial, negative actions. To the extent that these conditions differ, our account that a change in the meaning of the initial information plays a crucial role in revision is supported.

### **Method**

**Participants.** Two hundred ninety-nine participants were recruited on Amazon's Mechanical Turk website (mturk.com) for this study in return for \$0.75 in compensation (50.3% male;  $M_{\text{age}} = 33.9$  years,  $SD = 11.7$ ). We intended to collect data from 300 participants, but a technical error prevented one participant from completing the study, and that person was still compensated; thus, the study received data from only 299 individuals. As in Experiment 1a, this number of participants allowed us to fill each of our between-participants story conditions with data from approximately 100 people.

**Materials.** The story was identical to that used in Experiment 1a, except with an additional between-participants story condition added which served as a positive control

condition. This was meant to present participants with a piece of information about Francis West at time 2 that would be equally as positive as the fire rescue, but not providing an explanation for his initial, seemingly negative behaviors. Thus, the positive control condition would associate strong positive information with Francis just as the fire rescue condition does, while not justifying the recasting of the prior negative information. This comparison could illuminate whether the effect observed in Experiments 1a and 1b is due to the addition of an extremely positive piece of information to the participant's corpus of knowledge about Francis West to such a degree that it "swamps out" the previously learned negative information without any revision of the prior associative expressions per se, or as we predict, depends on a reinterpretation of the prior information.

In the positive control condition at time 2 participants read the following statement, pretested in a separate sample to be equally as positive to the action of saving two children from a raging fire:<sup>9</sup> "At a different point in time, Francis West was in the news because he was at a subway station when he noticed that a baby had crawled and fallen onto the tracks below. Seeing a rapidly approaching train, Francis jumped down onto the tracks, grabbed the baby, and climbed up to safety a split-second before the train came roaring past." Besides this difference, the new positive control condition proceeded in an identical fashion to the other two.

**Procedure.** The procedure was the same as in Experiment 1a, except for the addition of the positive control condition, and the manipulation check asking participants to identify the final

---

<sup>9</sup> Fifty Mturk workers read eight heroic actions that a hypothetical individual might do (e.g., "Waded into the water above Niagara Falls to save two stranded kids") along with a description similar to the fire rescue condition ("Ran into two burning homes to save two children from a fire"). These nine actions were presented in random order on a single screen, and participants evaluated how positively they would view a person who did the action on a 1-100 scale (Not positive at all to As positive as possible). The fire rescue behavior ( $M = 92.4$ ,  $SD = 10.5$ ) and the subway rescue scenario ( $M = 91.7$ ,  $SD = 16.1$ ) did not differ,  $t(49) = .54$ ,  $p = .595$ .

information they read about Francis, with 3 answer choices reflecting the time 2 information presented in the three story conditions. Besides these changes, participants viewed the same stories and took the same AMPs and questionnaire measures as used in Experiment 1a. At the end of the study, participants completed measures for an unrelated investigation.

## Results

**Data preparation.** In line with Payne et al. (2005), 14 participants familiar with Mandarin or Cantonese (4.7% of cases), and 17 participants who used a single key on every trial of at least one of the two AMPs (5.7% of cases), were excluded, leaving 268 cases for analysis.

**Implicit evaluations toward Francis West.** We assessed implicit evaluations toward Francis West by analyzing judgments of ideographs in a 2 (Measurement Time: time 1 and time 2) x 2 (Prime Person: Francis West and neutral) x 3 (Story Condition: control, fire rescue, or subway rescue) mixed design, with the first two factors manipulated within-participants and the third manipulated between-participants.

Every effect in the design was statistically significant, but of most interest, the three-way interaction between time, prime person, and story condition obtained,  $F(2,265) = 20.004, p < .001, \eta_p^2 = .131$ . There was once again no interaction between time and prime person in the control condition,  $F(1,265) = .16, p = .692$ , with only a main effect of prime person,  $F(1,265) = 57.17, p < .001, \eta_p^2 = .177$ , such that Francis West was evaluated less positively ( $M = .40, SD = .26$ ) than neutral faces ( $M = .64, SD = .20$ ).

Once more, in the fire rescue condition there was a significant interaction between time and prime person,  $F(1,265) = 76.87, p < .001, \eta_p^2 = .225$ . At time 1, Francis was significantly less implicitly positive ( $M = .39, SD = .27$ ) than the neutral faces ( $M = .64, SD = .19$ ),  $F(1,265) =$

41.43,  $p < .001$ ,  $\eta_p^2 = .135$ . However, at time 2, Francis was significantly more positive ( $M = .67$ ,  $SD = .23$ ) than the neutral faces ( $M = .53$ ,  $SD = .21$ ),  $F(1,265) = 14.78$ ,  $p < .001$ ,  $\eta_p^2 = .053$ .

In the subway condition, there was also a significant interaction between time and prime face,  $F(1,265) = 12.65$ ,  $p < .001$ ,  $\eta_p^2 = .046$ . At time 1, Francis was significantly less implicitly positive ( $M = .36$ ,  $SD = .27$ ) than the neutral faces ( $M = .64$ ,  $SD = .21$ ),  $F(1,265) = 54.83$ ,  $p < .001$ ,  $\eta_p^2 = .171$ . At time 2, Francis was still significantly less positive ( $M = .50$ ,  $SD = .30$ ) than the neutral faces ( $M = .62$ ,  $SD = .24$ ),  $F(1,265) = 11.45$ ,  $p = .001$ ,  $\eta_p^2 = .041$ . The increase in implicit positivity of Francis West from time 1 to time 2 was significant,  $F(1,265) = 17.74$ ,  $p < .001$ ,  $\eta_p^2 = .063$ . Thus, though the subway rescue condition significantly attenuated the implicit negativity of Francis relative to neutral faces, only the fire rescue condition evidenced a significant revision of a negative evaluation into a positive one. Figure 3 displays the mean proportion of ideographs judged more pleasant than average at time 2 for both Francis and the neutral faces within each story condition.

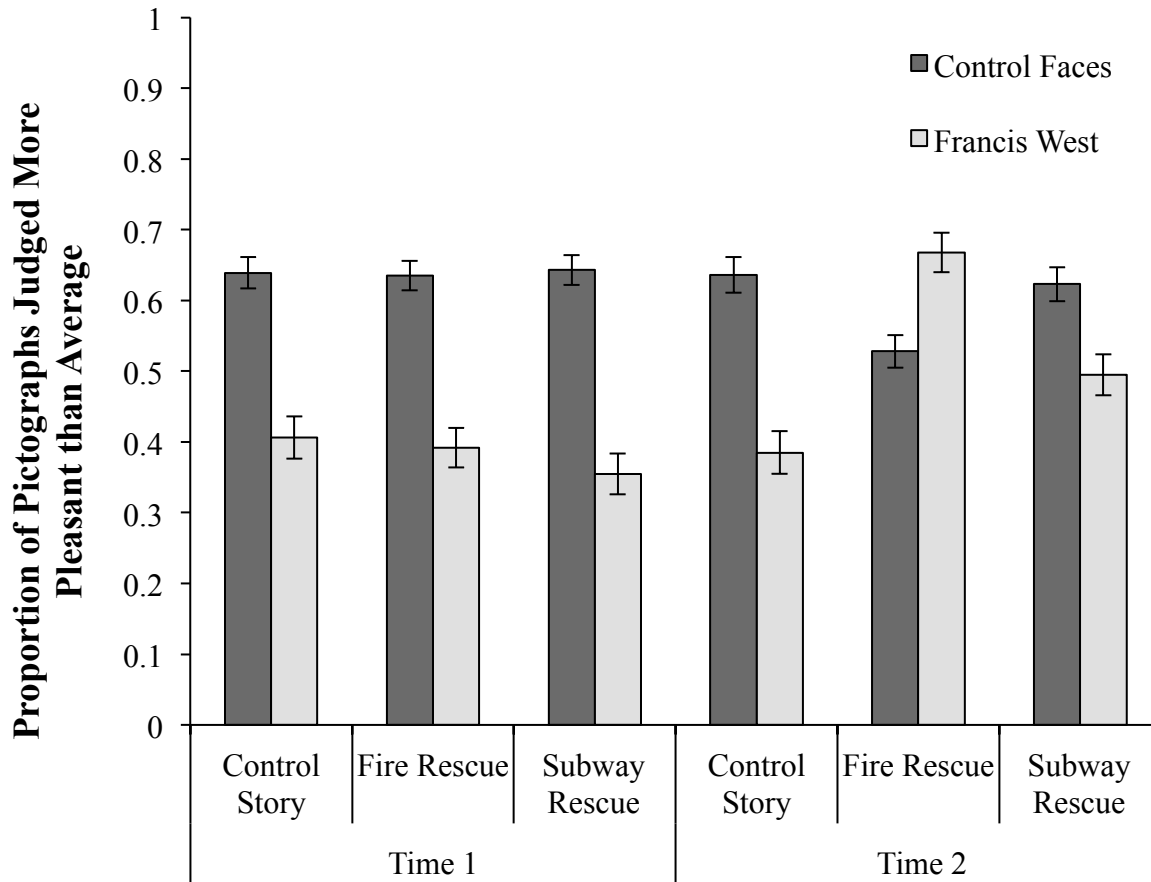


Figure 3. Mean proportion of ideographs judged more pleasant than average in Experiment 2, by measurement time, story condition, and face prime. Error bars are standard errors.

**Explicit evaluations toward Francis West.** The effect of story condition on explicit evaluations toward Francis West at time 2 was assessed using a one-way ANOVA (control, fire rescue, or subway rescue). As before, explicit evaluations toward Francis West at time 2 varied between conditions,  $F(2,265) = 430.42, p < .001, \eta_p^2 = .765$ . Participants explicitly liked Francis the most in the fire rescue condition ( $M = 5.86, SD = 1.35$ ), followed by the subway rescue condition ( $M = 2.45, SD = 1.18$ ), and finally the control condition ( $M = 1.21, SD = 0.60$ ). Each of the simple comparisons between conditions was significant, all  $ps < .001$ . In addition, one-sample t-tests revealed that each group mean significantly diverged from the midpoint, all  $ps < .001$ .

.001; thus, as with implicit evaluations, the subway rescue information did not make Francis West explicitly positive, though he was seen as less negative than in the control condition.

**Story condition effects on other questionnaire measures.** Subjective confusion with the story was lowest in the fire rescue condition ( $M = 2.01$ ,  $SD = 1.58$ ), higher in the control condition ( $M = 3.15$ ,  $SD = 1.95$ ), and highest in the subway rescue condition ( $M = 3.62$ ,  $SD = 1.96$ ),  $F(2,265) = 18.85$ ,  $p < .001$ ,  $\eta_p^2 = .125$ . Including subjective confusion in the analysis of implicit evaluations did not produce any interactions or change the significance of any effects.

As in Experiment 1a, there was also a story condition effect on the reported extent to which participants thought about the new information that they were presented with at time 2,  $F(2,265) = 3.03$ ,  $p = .05$ ,  $\eta_p^2 = .022$ . Participants thought the most about the information in the fire rescue condition ( $M = 5.96$ ,  $SD = 1.40$ ), and less in both the control condition ( $M = 5.56$ ,  $SD = 1.34$ ), and the subway rescue condition ( $M = 5.49$ ,  $SD = 1.42$ ). Extent of thinking was greater in the fire rescue condition relative to the others,  $F(1,265) = 5.93$ ,  $p = .016$ ,  $\eta_p^2 = .022$ , but the subway rescue and control conditions did not differ,  $F(1,265) = .10$ ,  $p = .758$ . This suggests that participants thought most about the information in the fire rescue condition.

**Moderation by extent of thinking.** Although the revision effect was not moderated by the self-reported “extent of thinking about the story” measure in Experiment 1a, when we added this as a covariate in the implicit evaluations analysis in the present experiment, the four-way interaction (between time, prime person, story condition, and extent of thinking) was marginal,  $F(2,262) = 2.74$ ,  $p = .066$ ,  $\eta_p^2 = .021$ . Examining the story conditions separately revealed that the time X prime person X extent of thought interaction was not significant in the control or subway rescue conditions, both  $ps > .5$ , but was in the fire rescue condition,  $F(1,93) = 6.14$ ,  $p = .015$ ,  $\eta_p^2 = .062$ . The effect was such that the time X prime person interaction (the revision effect) was

stronger at high levels of thinking about the story (+1 SD) ( $F[1,93] = 55.00, p < .001, \eta_p^2 = .372$ ) than at low levels (-1 SD) ( $F[1,93] = 12.03, p = .001, \eta_p^2 = .115$ ), though as the numbers show, the revision effect was still significant with less reported thinking.

## Discussion

Building on the initial studies, the control positive condition appeared markedly different from the fire rescue condition. Whereas reading about the fire rescue produced significant reversal, reading about the subway story reduced, but did not eliminate, participants' initial, negative impression. These results are consistent with our account that the ability of new information to prompt reinterpretation of prior information is key to full revision in this paradigm. The reduction, but not reversal, of negative implicit evaluations in the subway condition is consistent with the idea that the subway information was simply added to, but did not reverse, the initially learned negative information. In this sense the subway control condition bears some similarity to learning in the Bob paradigm, in which counter-attitudinal learning proceeds by presenting participants with unconnected statements about Bob that are opposite in valence, but do not contradict the previously learned information in any other way; in this paradigm, implicit evaluation revision generally proceeds slowly (e.g., Rydell & McConnell, 2006; Rydell et al., 2007) unless the initial impression is positive, and the new behavior is extreme and negative (Cone & Ferguson, in press). Although Cone and Ferguson (in press) found that a single, sufficiently extreme negative behavior was enough to significantly alter initial positive implicit evaluations toward a novel target, they too found that a single, extreme positive behavior was *not* enough to overturn an initial negative implicit impression. Together with the findings of the current study, this suggests that revision might be especially effective if



new information is added in such a way that also changes the meaning of the original information.

These findings also speak to some theoretical assumptions about how propositional information might modify implicit evaluations. The APE model (Gawronski & Bodenhausen, 2006) assumes that the affirmation of new, counter-attitudinal propositional information can create a new counter-attitudinal association, which could then affect implicit evaluations. However, here, participants in both the fire and subway conditions presumably affirmed (i.e., believed, processed) the new information, but reversal only happened in the former and not the latter case. This finding thus illustrates the (theoretical) importance of identifying *when* new propositional learning can modify implicit evaluations.

We still, however, do not know much about the process of reinterpretation. Are participants effortfully reinterpreting the previous information? Though self-reported extent of thinking marginally moderated the overall revision effect, the phrasing of this single question (“When you read that final piece of information about Francis West, to what extent did you think about the details you read earlier in the study?”) made it specific to how much the participants felt that they explicitly revisited the previous details. As such, it likely did not adequately tap other forms of thinking, such as extent of thinking about the *new* information or rapid comprehension that did not require deliberate revisiting of the old information. The item also does not indicate what participants did with the old information when revisiting it (reinterpret, rehearse, reject, etc.). We therefore take this marginal effect as only suggestive evidence that some type of active thinking is involved in revision in these studies, and in the following studies we examine which aspects of active thinking are most important in producing revision.

In our next study, we tested whether some minimal degree of effortful processing is required to produce revision of implicit evaluations. To do so, we examine if this revision requires effort, and therefore will be reduced when one is under high cognitive load.

### **Experiment 3**

#### **Method**

**Participants.** Four-hundred fifty-one individuals were recruited from Mturk to participate in exchange for \$1.00 (47.9% male;  $M_{\text{age}} = 34.86$ ,  $SD = 11.8$ ). A priori, we wanted to collect data from 450 participants so as to fill each of six between-participants conditions with 75 participants each. Given the results of our previous experiments, we determined via power analysis that a full 100 per cell was not necessary to achieve our effects, and reduced this amount to 75 for reasons of cost (while remaining above 90% power). Data from an additional participant were collected because one person completed the study without submitting the request for payment on Mturk.

**Materials.** The story used in Experiment 1a was used, and the stimuli were identical to those used in the previous studies. From the questionnaire, we dropped the item gauging difficulty making sense of the story after learning the final information (due to redundancy with the general confusion question) and the item asking the extent to which participants thought about the prior story details when learning the final information. That item had seemed to affect the revision results in one of the two studies in which it had been included (marginal interaction in Experiment 2) but not the other (Experiment 1a). We suspected that given the unburdened ability of all participants to consider the final information for as long as they wished, this “extent of thinking” measure might not effectively tap into meaningful variation in effortful thinking. To get more specifically at the type of thinking that might matter to the revision effect, we used the

following two items, solely in the low- and high-cognitive load conditions: One asking participants how much they went back to think about the *new* story details after they no longer had to remember the number, and a second item asking the same about the *old* story details, both on a scale from 1 (*Not at all*) to 7 (*Very much*). We did not ask participants to self-report their extent of thinking in the no-load conditions, as we chose to focus here on which type of information participants would selectively choose to mentally revisit once given the opportunity to after the relieving of cognitive load.

**Procedure.** All participants completed the first story session and the first AMP, and were then told that they would be reading one final piece of information about Francis West. Participants in the no load condition were then presented with this information and moved on to the second AMP as was done in Experiments 1-2. In the two cognitive load conditions, however, participants were informed that they would do an additional task while considering this information: they would need to maintain a number in memory, to be reported immediately after moving on from the new story information. Such a cognitive load induction technique has been used successfully in prior research (see, e.g., Gilbert & Osborne, 1989). When they understood these directions and were ready to proceed, participants were presented with either a random 2-digit number (low load) or random 8-digit number (high load) for twenty seconds. Then, the page automatically advanced to the final information about Francis West (either fire rescue or control). When participants were ready to advance, they were then presented with a textbox in which they had up to fifteen seconds to enter the number they were previously presented with. Following this, the page automatically advanced to the second AMP. All participants then completed the explicit questionnaire items, and finally an unrelated experiment.

## **Results**

**Data preparation.** Following the same procedure as our previous studies, 6 participants familiar with Mandarin or Cantonese were excluded from all analyses (1.3% of cases), as were 34 additional participants who used only one of the two response keys on all of the trials of at least one of the two AMPs (7.5% of cases), following established procedure (Payne et al., 2005). This left 411 cases for analysis.

**Memory for the number in the low and high load conditions.** We coded the numbers that participants recalled in the low and high load conditions for errors (omitted or extra digits, digits out of order). Predictably, perfect recall of the number occurred more frequently in the low load (99.25%) than high load condition (80.42%),  $\chi^2(1) = 26.18, p < .001$ . To ensure that all data from the low and high load conditions are drawn from participants who had engaged in the cognitive load task (and thus experienced the manipulation as intended), we included in all analyses only those participants who perfectly recalled the number (Gilbert & Osborne, 1989). Getting the number correct had no effect on comprehension score,  $F(1,407) = .072, p = .788$ , nor did its interaction with story condition,  $F(1,407) = .064, p = .800$ . Where instructive, we also report findings for those participants who failed to correctly recall the number assigned to them.

**Implicit evaluations toward Francis West.** We submitted the proportion of ideographs judged as more pleasant than average to a 2 (Measurement Time: time 1 and time 2) x 2 (Prime Person: Francis West and Neutral) x 2 (Story Condition: Control or Fire Rescue) x 3 (Cognitive Load: No Load, Low Load, or High Load) mixed ANOVA, with the first two factors manipulated within-participants and the latter two manipulated between-participants.

Replicating the previous studies, the three-way interaction between measurement time, person, and story condition was significant,  $F(1,376) = 36.55, p < .001, \eta_p^2 = .089$ . However, this effect was significantly moderated by cognitive load,  $F(1,376) = 3.43, p = .034, \eta_p^2 = .018$ .

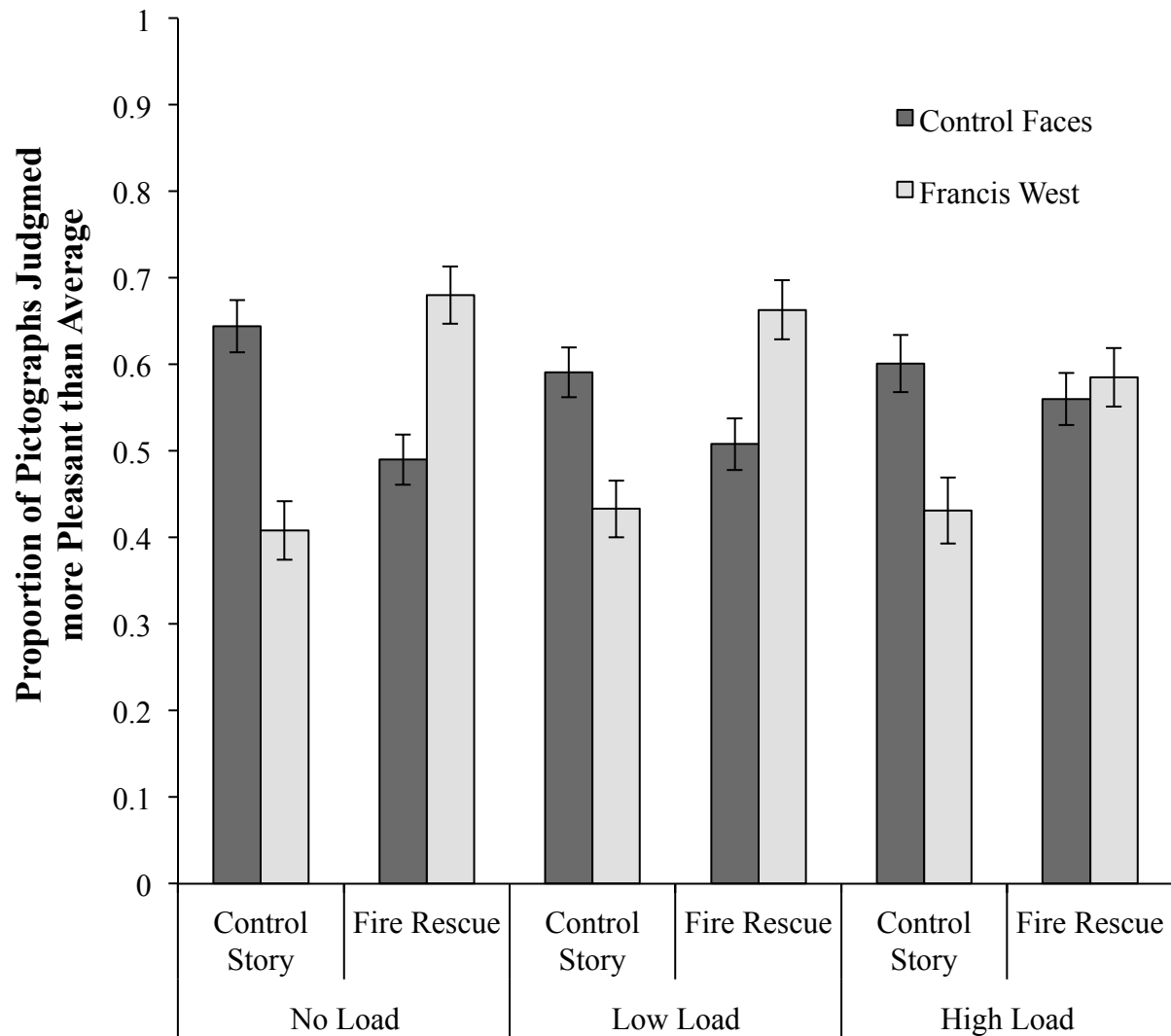
Planned follow up analyses indicated that the contrast of high load vs. the other conditions (no load and low load) moderated the revision effect,  $F(1,376) = 6.39, p = .012, \eta_p^2 = .017$ , while the comparison of no load vs. low load did not,  $F(1,376) = .46, p = .498$ . In the high load condition, there was no significant interaction between time, person, and story,  $F(1,376) = 1.70, p = .193$ , whereas in the other two conditions there was,  $F(1,376) = 43.75, p < .001, \eta_p^2 = .104$ . As a result, in the high load condition Francis West did not exceed the neutral faces in positivity at time 2 in the fire rescue condition,  $F(1,376) = .245, p = .621$ , while he did in the other two load conditions combined,  $F(1,376) = 25.102, p < .001, \eta_p^2 = .063$ . Figure 4 displays the mean proportion of ideographs judged pleasant for both Francis West and the neutral faces within each story and cognitive load condition at time 2.<sup>10</sup> The interaction between the revision effect and the high load vs. other conditions contrast remains significant even if including only those participants who had perfect comprehension, correctly identified Francis in the photo lineup, and/or correctly identified the final story detail they had been presented with, all  $ps < .05$ .

Although the exclusion of those not perfectly remembering the number during the cognitive load task reduces the sample size by 28 individuals in the high load condition to 115 (80.42%), this is not likely to be responsible for the lack of a time X person X story effect in the high load condition. This is because when we conducted our revision analysis on solely those 28 participants in the high load condition who failed to recall the number, we found significant revision among this sample, evidenced by a significant interaction between time, person, and story,  $F(1,26) = 6.78, p = .015, \eta_p^2 = .207$ . If the 28 participants who failed to recall the number

---

<sup>10</sup> The lack of interaction between time, person prime, and story in the high load condition is not due to differences at time 1. At time 1, there was no interaction between person, story condition, and cognitive load, whether the latter is coded as the original three-level factor,  $F(2,376) = .71, p = .492$ , or a contrast of the high load group vs. the other two: no load and low load;  $F(1,378) = 1.33, p = .250$ . Indeed, in the high-load condition at time 1, Francis was more negative than neutral faces in both the control and fire rescue conditions (both  $ps < .001$ ).

in the high load condition show significant revision, then reduced sample size is unlikely to be the explanation for why the 115 participants who did remember the number failed to show it.



*Figure 4.* Mean proportion of ideographs judged more pleasant than average in Experiment 3 at time 2 by cognitive load condition, story condition, and face prime. Error bars are standard errors.

***Moderation by extent of thinking.*** We conducted two additional analyses to determine whether, in the load conditions, the extent to which participants reported that they went back to consider the new information and the old information moderated the size of the revision effect.

In the first analysis, we examined the moderating effect of thinking about the new information. The three-way effect of time, person, and story condition was moderated by the amount of thinking about the new information,  $F(1,244) = 4.27, p = .04, \eta_p^2 = .017$ . Probing this interaction 1 SD above and below the mean level of thinking about the new information, we found that the revision effect (time X person X story condition) was significant at high levels of going back to think about the new information in the story,  $F(1,244) = 15.76, p < .001, \eta_p^2 = .061$ , but not at low levels of going back to think about the new information,  $F(1,244) = .89, p = .347$ . Adding cognitive load condition to this model (low vs. high load) did not qualify the effect,  $F(1,240) = .12, p = .731$ . Thus, the experience of going back to think on the new information appears to moderate the revision effect, and this was the case regardless of the amount of cognitive load.

In the second analysis, we examined the moderating effect of thinking about the old information. The three-way revision effect was not qualified by the amount of thinking about the old information,  $F(1,244) = .03, p = .854$ . However, when cognitive load was added to the model, there was a significant interaction between time, person, story condition, cognitive load, and extent of thinking about the old information,  $F(1,240) = 9.52, p = .002, \eta_p^2 = .038$ . Investigating this interaction revealed that for low levels of thinking about the old information, cognitive load did not moderate the revision effect, which was strong. However, at high levels of thinking about the old information, cognitive load moderated the revision effect,  $F(1,240) = 12.403, p = .001, \eta_p^2 = .049$ . This effect is such that in the fire rescue condition, there is reversal only for low load ( $F(1,240) = 19.103, p < .001, \eta_p^2 = .074$ ), but not high load ( $F(1,240) = .608, p = .436$ ). In high load, Francis is only neutral at time 2, with unfamiliar faces and Francis not differing,  $F(1,240) = .989, p = .321$ , even though he's significantly negative at time 1. Thus, when participants reported thinking a lot about the *old* information after no longer needing to

remember the number, this tended to worsen the revision effect in the high cognitive load condition. Although one might have plausibly supposed that such thinking indicates active efforts at reinterpretation, it seems that such thinking is not an indicator of successfully making sense of the story. If thinking about the new information but not the old information seems to be linked with strong revision, this may mean that changing one's interpretation of the story is experienced as extensive thought about the *new* information, but not the old.

**Explicit evaluations toward Francis West.** We assessed explicit liking of Francis West in a 2 (Story Condition: Control vs. Fire Rescue) x 2 (Cognitive Load: No Load, Low Load, or High Load) ANOVA. Results indicated a main effect of story condition,  $F(1,376) = 1281.76, p < .001, \eta_p^2 = .773$ , such that liking was higher in the fire rescue condition ( $M = 5.69, SD = 1.59$ ) than in the control condition ( $M = 1.25, SD = 0.61$ ). There was no main effect of cognitive load condition,  $F(2,376) = 1.66, p = .192$ , or interaction between story condition and cognitive load condition,  $F(2,376) = 2.17, p = .116$ . Thus, while cognitive load affected implicit evaluations, it did not affect explicit evaluations toward Francis.

## **Discussion**

Consistent with our account that active thinking about the information is central to our reinterpretation effects, these results show that the availability of at least minimal cognitive resources is a necessary condition for full revision. Those in the high load condition, in comparison with those in low load or no load conditions, showed no moderation of their implicit evaluations by condition. Participants with less or no load, meanwhile, showed the pattern of results observed in our previous studies: significant reversal in the fire rescue condition and no change in the control condition. The failure of those in the high load condition to show the revision effect was not attributable to a failure to learn the new story information, or even an



inability to come to the correct conclusions about the story when answering the comprehension questions. Instead, it seems that the high load burden interrupted not the *ability* to process the new information, which participants were able to do when queried, but rather their *tendency* to do so. The moderation of the revision effect by extent of going back to think about the new information suggests that among those who did, the reversal effect occurred. However, the failure of revision overall in the high load condition does suggest that on the whole, participants in this condition moved on to the AMP without doing this.

One interesting question is why did the self-report measure of going back to think about the new information produce a stronger reversal effect, while thinking about the old information seemed to do the opposite (at least under high cognitive load)? This question raises the issue of the type of active thinking required for revision in our studies. It is hard to interpret what participants might mean when they report thinking about the “old” versus the “new” information. These questions are too broad to pinpoint whether participants are rehearsing, reinterpreting, elaborating, rejecting, etc. Thus, in the next study we more precisely measured the type of thinking that we predict should produce revision: recognition that the new information produced *reinterpretation* of what had been previously learned about Francis.

#### **Experiment 4**

The comparison of the fire rescue and subway rescue conditions in Experiment 2 suggested that the revision effect emerged because of reinterpretation. In other words, revision occurred in the fire but not in the subway condition because of the degree to which the new, counter-attitudinal information was able to *explain away* the initial information only in the former. The first goal of Experiment 4 was to extend the evidence in support of this argument by measuring the degree to which participants report reinterpreting the earlier story information

after learning the fire rescue information about Francis West, to show that such subjective reinterpretation does indeed predict the revised implicit evaluations of Francis.

A second goal of Experiment 4, however, was to begin to address how reinterpretation might relate to, or differ from, other forms of thinking about the new information that could be responsible for our effects. In particular, to this point it is unclear how reinterpretation relates to research on elaboration. Much research has suggested that the extent to which one thoughtfully processes new information is an important predictor of its effect on explicit evaluations (for reviews, see Barden & Tormala, 2014; Petty, Haugtvedt, & Smith, 1995). When persuasive information is more thoroughly elaborated, it has a more powerful impact on impressions. A few studies have found similar effects of elaboration on established (Briñol, Petty, & McCaslin, 2009; Horcajo, Briñol, & Petty, 2010) or novel (Smith, De Houwer, & Nosek, 2013; Wyer, 2010) implicit evaluations. Of most relevance to our present purposes, Wyer (2010) found in one study that new information that suggested a reinterpretation of prior details did not produce revision of implicit evaluations *unless* participants were able to revisit all of the prior information upon which they had based their first impressions. Wyer (2010) argued, in line with Petty et al. (2006), that this suggested that for the old evaluation to be effectively tagged as false such that it would no longer impact implicit evaluations, participants needed this opportunity to carefully elaborate on the new revelation.

Is the reinterpretation in our studies effective because it forces participants to engage in extensive amounts of elaboration on the new information, but not necessarily reinterpretation in particular? That is, perhaps the critical ingredient in reinterpretation in our paradigm is that it simply is an effective way to get people to do a lot of thinking about the new information, regardless of whether those thoughts are specifically about reinterpreting the meaning of Francis’

earlier actions. From this perspective, the reason we have not seen revision in the subway condition (Experiment 2), for instance, is because that information, for whatever reason, was not sufficiently surprising or interesting to trigger enough elaboration (on that new information) to produce revision.<sup>11</sup> And, the findings from Experiment 3 – that high cognitive load prevented full revision of implicit evaluations – show that some amount of active thinking is necessary to produce revision, but do not disambiguate what sort of elaboration participants are engaging in (reinterpreting the earlier details or otherwise).

If the critical aspect of our paradigm is that it produces a large amount of general elaboration on the new information, such that change is driven by the degree to which people think carefully about the new information in general rather than reinterpretation of the earlier details in particular, then just the degree of thinking about the new information should predict revision, and any self-reported reinterpretation would not independently contribute to the effect. However, if reinterpretation is the *specific form* of elaborative thinking that drives the effects, then we should find that the belief that new information changes the meaning of the old information predicts revision, even when the extent of thinking more generally is controlled. Such a finding would imply that the proximal mechanism in our studies is the recognition that the new information changes the meaning of the old. We test these two accounts in the next study.

---

<sup>11</sup> Indeed, we found that participants in the subway rescue condition in Study 2 reported thinking about the prior story details less than those in the fire rescue condition. However, when we went back and compared the fire rescue and subway rescue conditions with only those participants selecting the highest value of “7” on the thought extent scale, story condition still moderated the interaction between time and person,  $F(1,71) = 11.66, p = .001, \eta_p^2 = .141$ . At time 2, Francis was still significantly more negative than control faces in the subway condition ( $F(1,71) = 5.29, p = .024, \eta_p^2 = .069$ ) but significantly more positive than control faces in the fire condition ( $F(1,71) = 14.52, p < .001, \eta_p^2 = .170$ ). So, reporting the max value on the measure of how much they thought about the story does not fully account for the difference in revision between the two conditions.

We included only the fire rescue condition, and added three items: an item about reinterpretation (how much does the new information change the meaning of the prior events), and two items gauging degree of thinking about the new information more generally (how rapidly vs. gradually one's thinking proceeded, and how extensively one deliberated about the new information). We predicted that degree of thinking carefully (either gradually and/or extensively) might significantly predict revision, in line with previous research (Briñol et al., 2009; Petty et al., 2006; Wyer, 2010). However, we also predicted that the reinterpretation item would also uniquely predict revision even while controlling for careful thinking. This would suggest that the reinterpretation happening in our studies – which is responsible for the revision effects – is a more specific mechanism than general elaboration on the new information.

## **Method**

**Participants.** We recruited 75 participants from Mechanical Turk to participate in the current study in exchange for \$1.75 (36% male;  $M_{\text{age}} = 36.56$  years,  $SD = 11.47$ ). This smaller sample size was determined a priori based on the lack of between-participants conditions in the current study.

**Materials.** To assess whether revised implicit evaluations of Francis West would be predicted by the degree to which participants reinterpreted the earlier story details, we asked participants to respond to the following question: “When you got the new information about Francis West a moment ago, how much did this new information change the meaning of Francis West's earlier actions?” on a scale from 1 (Not at all) to 9 (A large amount).

To begin to address whether reinterpretation in particular is predictive of revised implicit evaluations of Francis West, rather than elaborative thinking more generally, we added two further questions. The first was designed to measure the sense participants had of how quickly or

gradually their thoughts about Francis came together after learning the new information. Specifically, they read: “Sometimes, our thoughts come together quickly. At other times, our thoughts come together more gradually. When you got the new information about Francis West a moment ago, did your thoughts about the meaning of Francis' actions come together quickly or more gradually?” Participants responded on a scale from 1 (*quickly*) to 9 (*gradually*). Our second question aimed at tapping into the degree to which participants elaborated more generally (vs. reinterpretation in particular) focused on how extensive participants felt their thinking to be. They read: “Sometimes, we deliberate a lot, and our thinking is very extensive. At other times, we deliberate less, and our thinking is less extensive. When you got the new information about Francis West a moment ago, how much thinking did you do - not much deliberation or a lot of deliberation?” and responded on a scale from 1 (*Not much deliberation*) to 9 (*A lot of deliberation*). All of the other explicit measures from Experiment 3 were included, except for those dealing with the cognitive load manipulation from that study (including the extent to which they went back to think about the new and old information after no longer needing to remember the number), since the cognitive load task was not included here.

**Procedure.** All participants completed the fire rescue condition. They first read the Time 1 information, followed by the first AMP, and then were presented with the Time 2 fire rescue information as presented in Experiments 1a, 2, and 3, with one alteration to their instructions: To encourage more variability in the extent of reinterpretation, we simply asked participants to think about the final information, rather than specifically to think about how it relates to what they had previously read.

Next, right after reading the Time 2 (fire rescue) information, but before the second AMP, participants responded to the three questions regarding the nature of their thoughts at the

time of learning the final details about Francis West. To increase the chance that participants would discriminate between these three (potentially highly related) questions, we presented all three questions on the same screen and required participants to read all of them in advance, for at least 30 seconds, before moving on to answer them. To reduce potential noise from different question orders, we fixed the order of the questions, such that the meaning change question came first, followed by the thought speed question, and finally the deliberation extent question. Participants then completed the second AMP, the rest of the explicit measures, and were thanked, debriefed, and compensated.

## Results

**Data preparation.** Following our procedure from the previous studies, we excluded from all analyzes the data from those participants who reported that they knew Mandarin or Cantonese (2 participants; 2.67%), and any additional participants who used a single response key on all trials of at least 1 AMP (4 participants; 5.33%). This left 69 participants in the analysis.

**Implicit evaluations toward Francis West.** Implicit evaluations were once again measured from the average proportion of pleasantness judgments of ideographs following the different face primes on the AMP. These AMP judgments were analyzed in a 2 (Measurement Time: time 1 and time 2) x 2 (Prime Person: Francis West and control faces) repeated-measures ANOVA. The anticipated interaction between measurement time and prime person was significant,  $F(1,68) = 23.33, p < .001, \eta_p^2 = .255$ . At time 1, Francis West was significantly less implicitly positive than control faces: Ideographs following Francis primes were judged to be more pleasant than average significantly less often ( $M = .37, SD = .28$ ) than ideographs following control face primes ( $M = .57, SD = .20$ ),  $F(1,68) = , p < .001, \eta_p^2 = .260$ . At time 2, however, implicit evaluations had reversed. Ideographs following Francis primes were judged to

be more pleasant than average significantly more often ( $M = .66$ ,  $SD = .27$ ) than ideographs following control faces primes ( $M = .53$ ,  $SD = .23$ ),  $F(1,68) = 9.30$ ,  $p = .003$ ,  $\eta_p^2 = .120$ .

***Subjective thought measures and implicit evaluations.*** Next, to examine the relationship between the type of thinking that participants reported doing when reading the final information about Francis West and revised implicit evaluations toward him, we conducted a planned multiple linear regression analysis. The dependent variable was the average proportion of “pleasant” judgments following Francis West primes at time 2 for each participant, with average pleasantness judgments of ideographs following Francis primes at time 1, and control face primes at time 1 and time 2 entered as three covariates. The three measures of thought type (extent to which the new information changed the meaning of the old, speed of thought, and extent of deliberation) were the key predictors. All six of the model predictors were entered simultaneously in a single step. This allowed us to examine the potential for each predictor to have independent influences on final implicit evaluations of Francis West.

Results showed that self-reported extent to which the new information changed the meaning of the prior details had a uniquely predictive relationship with final implicit positivity of Francis West,  $\beta = .287$ ,  $t(62) = 2.30$ ,  $p = .025$ . However, the measure of whether participants felt their thoughts came together rapidly vs. gradually had no relationship with time 2 implicit evaluations of Francis,  $\beta = .001$ ,  $t(62) = .009$ ,  $p = .993$ , and neither did the measure of how extensive participants reported their thinking to be,  $\beta = .052$ ,  $t(62) = .420$ ,  $p = .676$ . To check the robustness of the relationship between the measure of meaning change and implicit evaluations of Francis, as well as to determine whether either of the other two thought measures would predict implicit evaluations if the other thought measures were omitted, we conducted a series of exploratory follow-up regressions. Specifically, we examined regressions that included all

possible subsets of the three thought measures (with the same covariates of other AMP trial types), and found that in none of these models did either the thought speed or deliberation extent measure produce a significant effect, all  $ps > .5$ . Additionally, the meaning change measure never became non-significant, all  $ps < .05$ .

The sample as a whole also strongly endorsed the view that the new information changed the meaning of the prior events ( $M = 8.59$  out of 9,  $SD = .99$ ), that their thinking proceeded quickly rather than gradually ( $M = 2.41$  out of 9,  $SD = 2.10$ , where higher values indicate more gradual thinking), and that their deliberation was not very extensive ( $M = 2.87$  out of 9,  $SD = 2.44$ ), suggesting that revision here tends to be experienced as relatively easy (provided that cognitive resources are not maximally strained as in Experiment 3). Indeed, the reported extent to which the meaning of the initial story had changed was negatively correlated with both the degree to which thinking proceeded gradually,  $r(67) = -.41, p < .001$ , and extent of deliberation,  $r(67) = -.36, p = .002$ . The degree to which thinking was gradual (vs. fast) correlated positively with the extent of deliberation,  $r(67) = .45, p < .001$ .

**Explicit evaluations toward Francis West.** We once again used an average of responses on the 6 questions gauging liking of Francis West to assess changes in explicit liking over time. In a paired-samples t-test we found that, unsurprisingly, explicit liking of Francis West was much higher at time 2 ( $M = 6.27, SD = .97$ ) than at time 1 ( $M = 1.21, SD = .51$ ),  $t(68) = 38.83, p < .001$ .

Further, we conducted a similar multiple regression analysis to that performed on implicit evaluations toward Francis West. The index of explicit liking at time 2 was regressed on liking at time 1, as well as the belief that the new information changed the meaning of the old, the gradualness of thought, and the extent of thought. Paralleling the results with implicit



evaluations, we found that the extent to which participants felt the new information changed the meaning of the old information predicted time 2 explicit liking of Francis,  $\beta = .664$ ,  $t(64) = 6.87$ ,  $p < .001$ . However, both the thought speed measure ( $\beta = -.178$ ,  $t[64] = -1.70$ ,  $p = .095$ ) and the deliberation extent measure ( $\beta = .070$ ,  $t[64] = .69$ ,  $p = .494$ ) did not.

## **Discussion**

The results supported our account that reinterpretation in the Francis West paradigm operates through a separate mechanism from general elaborative thinking. Although greater contemplative thought (in terms of either self-reported thought speed or extensiveness of deliberation) did not correlate significantly with greater revision, there was a unique, strong impact of belief that the new information changed the meaning of the prior story. Additionally, the distributions of responses on the measures and their negative correlation suggested that recognition of the new information's explanatory value was linked with having thoughts come together quickly rather than gradually, and with less extensive thinking. Reinterpretation seems to require at least a brief revisit of the prior story details so as to reframe their meaning, which requires the availability of at least some cognitive resources (Experiment 3), but it is not akin simply to extensive, general elaboration.

## **Experiment 5**

Experiment 2 demonstrated that information that reinterprets the prior events produced much stronger change (indeed, a reversal) than equally positive information that does *not* reinterpret the prior events. Building the case that reinterpretation is the operative mechanism in driving the revision effect in the fire rescue condition, Experiment 4 showed that the extent to which participants believed that the final information altered the meaning of the previous events was significantly correlated with final implicit evaluations of Francis West, while more general

measures of elaboration were not. However, we have not yet demonstrated that reinterpretation *per se mediates* the greater revision in the fire rescue condition relative to other conditions.

In this next experiment, we tested mediation by including solely the fire rescue and subway rescue control conditions from Experiment 2. In both conditions, we expected that the extent to which participants reported the new information to alter the meaning of the prior story events should predict the degree of their revision (some reinterpretation might occur among participants in the subway condition if they, say, suspected that his heroics might suggest that there was some unknown good reason behind his seemingly negative actions in his neighbors' homes). And, because the fire rescue information was expected to prompt this reinterpretation to a larger degree than the subway rescue information, we predicted that reinterpretation would mediate the effect of story condition on implicit evaluations. In addition, we retained the more direct measure of general elaboration from Experiment 4 (extent of thinking) to demonstrate that only reinterpretation, and not elaborative thinking more generally, would mediate the effect of story condition on revised implicit evaluations of Francis West.

## **Method**

**Participants.** Two-hundred ninety-six participants recruited on Mturk participated in return for \$1.75 (49% male;  $M_{\text{age}} = 33.72$  years,  $SD = 10.58$ ). We intended to recruit 300 participants (150 per between-participants condition), but a transient server error interrupted the experiment for 4 individuals, making it impossible for them to continue the study. They were compensated for their time, but their incomplete data were not included in any analyses.

**Materials.** Participants viewed the events from the same initial story used in Experiments 1a, 2, 3, and 4. The time 2 information consisted of that from either the fire rescue or subway rescue stories conditions used in Experiment 2. To measure the extent to which participants

subjectively reinterpreted the earlier story events in light of the new information, we asked participants to respond to the following single item adapted from Experiment 4: “When you got the new information about Francis West a moment ago, how much did this new information change the meaning of Francis West's earlier actions?” on a scale from 1 (Not at all) to 9 (Completely). To measure extent of more general elaborative thinking, participants responded to the same deliberation question from Experiment 4. All of the other explicit measures from Experiment 4 were included in this study, except for the item that gauged the speed with which thoughts came to mind.

**Procedure.** Participants were assigned to either the fire rescue or subway rescue condition, and completed the study in an identical fashion to Experiment 4. Immediately after reading the Time 2 information, participants responded to the subjective meaning change measure and the deliberation extensiveness measure before moving on to the second AMP. The order of these two questions was counterbalanced. To reduce the chance that the order in which the two questions were asked might influence participant responses, we again presented the questions on the same screen and asked participants to read both for at least 20 seconds before answering them. (The order of the two questions on the screen produced no significant effects in any analyses and is thus not discussed further.)

After answering the two questions about their thoughts when presented with the fire or subway rescue information, participants completed the second AMP, the rest of the explicit measures, and were thanked, debriefed and compensated.

## **Results**

**Data preparation.** In keeping with the exclusion criteria from our previous studies, we dropped all data from 9 participants for familiarity with Mandarin and/or Cantonese (3.0%) and

21 more for using a single key on every trial of at least one of the two AMPs, thus failing to follow instructions (7.1%). This left 266 cases for analysis.

**Implicit evaluations toward Francis West.** We analyzed average pleasantness judgments of ideographs on the AMP in a 2 (Measurement Time: time 1 and time 2) x 2 (Prime Person: Francis West and control faces) x 2 (Story Condition: fire rescue or subway rescue) mixed ANOVA, with the first two factors varying within-participants and the third between-participants. The anticipated interaction between time, prime person, and story condition obtained,  $F(1,264) = 44.41, p < .001, \eta_p^2 = .144$ . Simple effects tests revealed that at time 1, ideographs following the image of Francis West were rated more negatively than those following control faces in both the fire rescue condition ( $M_{\text{Francis}} = .38, SD_{\text{Francis}} = .28; M_{\text{Control}} = .61, SD_{\text{Control}} = .20; F[1,264], = 56.69, p < .001, \eta_p^2 = .177$ ) and the subway rescue condition ( $M_{\text{Francis}} = .36, SD_{\text{Francis}} = .28; M_{\text{Control}} = .65, SD_{\text{Control}} = .23; F[1,264], = 75.84, p < .001, \eta_p^2 = .223$ ). At time 2 in the fire rescue condition, implicit evaluations toward Francis West had reversed: ideographs following Francis primes were rated significantly more positively ( $M = .70, SD = .26$ ) than those following control primes ( $M = .51, SD = .24$ ),  $F(1,264) = 27.52, p < .001, \eta_p^2 = .094$ . However, at time 2 in the subway rescue condition, the initial implicit evaluations were attenuated but not reversed. Ideographs following Francis primes were still rated significantly more negatively ( $M = .41, SD = .28$ ) than ideographs following control face primes ( $M = .66, SD = .25$ ),  $F(1,264) = 48.35, p < .001, \eta_p^2 = .155$ . The interaction between time and prime person was significant in the fire rescue condition,  $F(1,264) = 108.58, p < .001, \eta_p^2 = .291$ , but not in the subway rescue condition,  $F(1,264) = .55, p = .459$ .

**Mediation by reinterpretation.** Next, we turned to our central interest in this experiment: Examining whether subjective degree of change in the meaning of the old story details in light of

the new information would uniquely mediate the fire rescue vs. subway rescue condition difference in final positivity toward Francis West, even when controlling for extent of general elaboration. Using the PROCESS tool for SPSS (Hayes, 2013), we conducted a bias-corrected bootstrap mediation analysis using 10,000 samples. The dependent variable was the proportion of ideographs judged pleasant following Francis West primes at time 2, with proportions following neutral primes at time 1 and 2 and Francis primes at time 1 entered as covariates. The independent variable was story condition (fire = 1, subway = 0) and the mediators were self-reported extent of change in the meaning of the initial story information and amount of deliberation. The two potential mediators were entered into a single model in parallel, but the interpretation of the results does not change in significance or direction if the mediators are run in separate analyses. Table 2 shows the zero-order correlations among all variables included in the model.

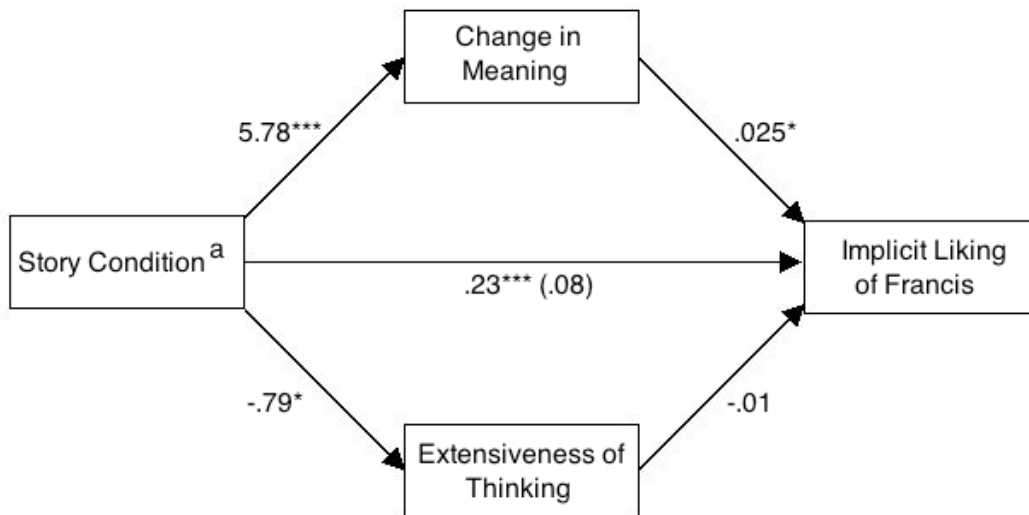
Table 2. *Zero-order correlations between all variables in the mediation model in Experiment 5*

Measure	1	2	3	4	5	6
1. Story condition						
2. Time 2 Francis pleasantness	.47***					
3. Time 2 control pleasantness	-.29***	-.36***				
4. Time 1 Francis pleasantness	.03	.29***	.04			
5. Time 1 control pleasantness	-.08	.01	.44***	-.08		
6. Meaning change	.89***	.49***	-.35***	.02	-.11	
7. Extent of deliberation	-.12†	-.07	-.07	-.07	.00	-.03

*Note.* Cell values are Pearson correlations. Story condition is coded 0 = subway rescue, 1 = fire rescue. Pleasantness covariates refer to the average portion of ideographs judged to be more pleasant than average, by time and prime type.

\*\*\*  $p < .001$ . †  $p < .1$ .

The analysis yielded a significant indirect effect for the mediation of story condition through reinterpretation, estimate = .146, 95% CI: [.0361, .2438], Sobel  $Z = 2.56$ ,  $p = .011$ . There was no parallel indirect effect through extent of deliberative thinking, estimate = .004, 95% CI: [-.0034, .0177], Sobel  $Z = .78$ ,  $p = .436$ . Figure 5 illustrates the mediation model.



*Figure 5.* Mediation of story condition effect on time 2 judgments of ideograph pleasantness following Francis West primes, through subjective change in meaning of the time 1 information and general extensiveness of thinking in Experiment 5. Slopes are unstandardized regression coefficients. Model controls for judgments of ideographs following Francis West primes at time 1 and control face primes at times 1 and 2.

<sup>a</sup> Coded 0: subway rescue, 1: fire rescue. \*\*\*  $p < .001$ , \*  $p < .05$ .

**Explicit evaluations toward Francis West.** We assessed changes in explicit evaluations of Francis West in a 2 (Measurement Time: time 1 and time 2) x 2 (Story Condition: fire rescue or subway rescue) mixed ANOVA, with measurement time varying within-participants and story condition varying between-participants. The dependent variable was the average of the six

explicit liking scales used in each of the prior experiments. Story condition significantly affected the change in explicit liking over time,  $F(1,264) = 826.91, p < .001, \eta_p^2 = .758$ . Francis increased in explicit positivity in the fire rescue condition from time 1 ( $M = 1.26, SD = .78$ ) to time 2 ( $M = 6.29, SD = 1.15$ ),  $F(1,264) = 2488.31, p < .001, \eta_p^2 = .904$ , as well as in the subway rescue condition from time 1 ( $M = 1.22, SD = .53$ ) to time 2 ( $M = 2.06, SD = 1.00$ ),  $F(1,264) = 62.43, p < .001, \eta_p^2 = .191$ , but was significantly more positive at time 2 in the fire rescue condition than the subway rescue condition,  $F(1,264) = 1017.61, p < .001, \eta_p^2 = .794$ . Furthermore, in a mediation analysis similar to that performed on implicit evaluations, we found that change in meaning of the initial story details mediated the effect of story condition on time 2 explicit liking of Francis West (controlling for time 1 explicit liking), estimate = 2.75, 95% bias-corrected bootstrap CI (10,000 samples): [2.1452, 3.2683], Sobel  $Z = 12.61, p < .001$ . On the other hand, there was no significant indirect effect through extent of deliberation, estimate = -.01, 95% bias-corrected bootstrap CI (10,000 samples): [-.0391, .0107], Sobel  $Z = -.45, p = .650$ .

## Discussion

As predicted, the extent to which participants reported that the new information changed the meaning of the earlier details of the story mediated the effect of the new information (fire rescue vs. subway rescue) on revised implicit evaluations of Francis West. That is, participants tended to engage in more reinterpretation in the fire rescue condition, and the extent to which they did so predicted the greater revision in that condition. Importantly, we also found no evidence that a more general measure of elaborative thinking – which asked participants to report whether they deliberated over the new information more or less extensively – mediated the condition difference in revision. Even when controlling for a potential indirect effect of general extent of thinking, the degree to which participants reported the meaning of the old information

was changed by the new information significantly mediated the difference between the fire rescue and subway conditions.

## **Experiment 6**

For our final study, we addressed a potential concern that the revised implicit evaluations produced in this work might not be durable, “real” change, but perhaps a transient effect in which the powerful new information is especially salient. This might produce a brief shift that masquerades as real change, only to revert back to the initial evaluation after the passing of time.

We thus sought to demonstrate the longevity of the revised implicit evaluations. Indeed, a changing temporal context has been noted as a potential source of spontaneous recovery of conditioned responses (Bouton, 1993; Bouton, Westbrook, Corcoran, & Maren, 2006), and so demonstrating no return of the initial negative implicit evaluation of Francis West in the revision condition would be informative. In showing the endurance of the revised implicit evaluations, we can therefore suggest that there is nontrivial durability and thus “realness” to these evaluations.

To examine the effects of time on implicit evaluation revision, we repeated our basic revision procedure. Then, participants were invited to return for a follow-up study three days later. At that time, they were told simply that they might remember reading a story about a man named Francis West, but that before they would be asked about this they were to complete a different task (a third implicit measure). To show the durability of revision, we expected to observe relatively no change in revised implicit evaluations across the delay. To keep things simple, we used only the fire rescue and control conditions.

## **Method**

**Participants.** Three hundred and one participants were recruited from Mechanical Turk to take part in this two-session study in return for \$1.50 paid compensation (53.5% male;  $M_{\text{age}} =$



32.02 years,  $SD = 10.08$ ). Because we were uncertain about how much attrition there would be between the two study sessions, and about whether the effect size would be much reduced after a delay, we opted a priori to collect data from 300 participants so as to fill each of two between-participants conditions with 150 participants. Data from an additional participant were recorded because one person completed the study without submitting the request for payment on Mturk.

**Materials.** The story materials and AMP stimuli used in this study for time 1 and time 2 were identical to those used in the previous experiments, including the control and fire rescue conditions, with one exception: At time 2, participants in that condition now read that he had a criminal history, yelled at children, and broke into the houses in revenge against them as well as to steal valuables (rather than that he started throwing rocks at the houses). This change was made to better equate the two conditions on the degree to which the final information provided motive for his actions. At time 3, the AMP used the same prime images as the preceding AMPs and one of two new randomly chosen sets of 40 ideographs as targets. Thus the same ideographs were never rated twice by an individual participant. The explicit evaluation toward Francis at time 3 was measured using the same scale as used at times 1 and 2. For exploratory purposes, we added a single item right before the demographic questions at time 2 asking participants to self-report their mood (“Indicate how you feel right now, that is, at the present moment”) on a scale from 1 (*very bad*) to 7 (*very good*), and an item gauging the extent to which they thought the story depicted “real” events (“To what extent do you believe that the Francis West story is based on real events?”) from 1 (*not at all*) to 7 (*completely*).

**Procedure.** After being assigned to either the fire rescue or control story conditions, participants completed the same procedure as in Experiment 3’s no cognitive load condition, without the questions about thought type. To minimize noise, all participants completed the

AMP prior to the explicit evaluation scale at each measurement instance. After completing the various questionnaire items at time 2 (explicit evaluation scale, comprehension checks, manipulation checks, mood, belief the story was real, and demographic questions), participants were then informed that they had the option of entering an email address so that we could contact them in three days with a short follow-up study, which they would receive extra compensation for completing. They were told that not doing so would have no impact on their compensation for what they had already completed. All but 6 did so. Approximately 3 days later, participants received an email inviting them back for the short final session of the study, and were given a window of 24 hours in which to complete it. 63.1% of participants returned and completed the final session. Attrition was equally likely in the two story conditions,  $\chi^2(1) = .12, p = .725$ .

Upon beginning the final session, all participants were told, “Three days ago you read a story about a man named Francis West. You will be asked to answer questions about him in a few moments; please do your best to answer these questions regardless of how much you remember. But first, there is another task to do.” They then completed the third AMP, followed by the third administration of the explicit evaluation scale (dubbed “time 3” hereafter).

## Results

**Data preparation.** Implicit evaluations on each of the three AMPs were computed in the same manner as done in previous studies, as were the three repetitions of the explicit evaluation scale. All analyses were conducted solely on those participants who completed the second session of the study. In addition, 10 participants were dropped for using one key on every trial of at least one AMP, thus disregarding instructions, and 1 more was dropped for familiarity with Mandarin or Cantonese. This left 179 cases for analysis.

**Implicit evaluations toward Francis West.** Implicit positivity on the AMP toward Francis West was assessed in a 2 (Story Condition: fire rescue or control) x 2 (Prime Person: Francis West vs. control faces) x 3 (Measurement Time: time 1, time 2, and time 3) Mixed ANOVA, with the first factor varying between-participants and the latter two within-participants. All effects were significant, including the crucial three-way interaction,  $F(2,175) = 25.60, p < .001, \eta_p^2 = .226$ . In Figure 6 we show the means and standard deviations of positivity toward the Francis and control primes in each of the story conditions and each of times 1, 2, and 3. Importantly, Francis West was significantly more negative than the control faces in the control condition at all measurement instances, as well as the fire rescue condition at time 1, all  $ps < .001$ ; additionally, he was more positive than the control faces in the fire rescue condition at times 2 and 3, both  $ps < .001$ . Thus, with only the barest of reminders about the study, implicit evaluations toward Francis persevered in both the fire rescue and control conditions (positive and negative, respectively) for the three days between session 1 and session 2 of the experiment. Time effects showed that in the control story condition, positivity toward Francis did not shift between any two measurement times, all  $ps > .1$ . In the fire rescue condition, Francis West was significantly more positive at time 2 ( $M = .70, SD = .23$ ) than at time 1 ( $M = .41, SD = .27$ ),  $p < .001$ , and was marginally less positive at time 3 ( $M = .66, SD = .23$ ) relative to time 2,  $p = .071$ . However, even at time 3 he was still more positive than at time 1,  $p < .001$ .

Consistent with prior testing, neither the degree of belief that the study depicted true events nor subjective mood moderated these results. Also, neither reduced the significance of the key interaction when added to the model.

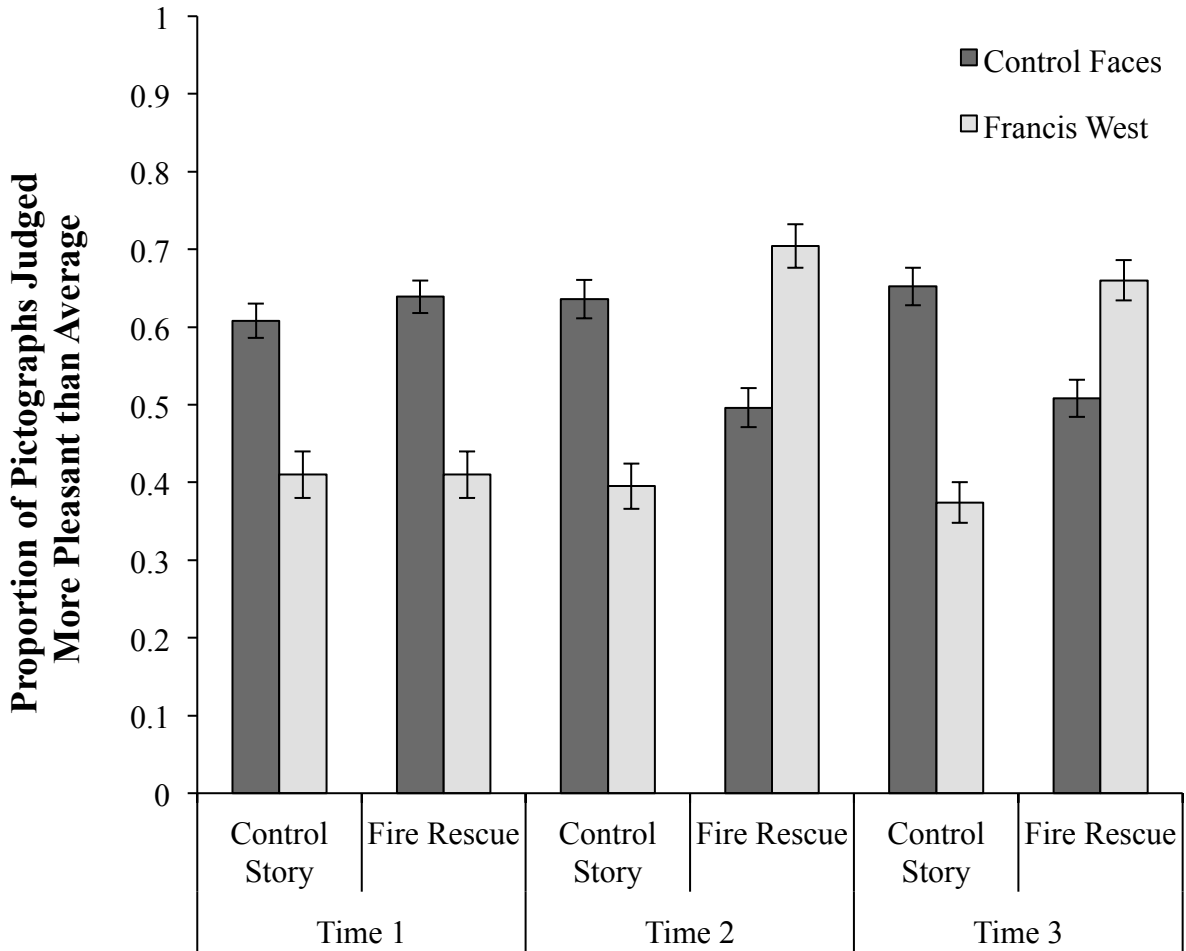


Figure 6. Mean proportion of ideographs rated as more pleasant than average in Experiment 6 by measurement time, story condition, and face prime. Error bars are standard errors.

**Explicit evaluations toward Francis West.** Explicit liking of Francis West was analyzed in a 2 (Story Condition: fire rescue or control) x 3 (Measurement Time: time 1, time 2, and time 3) mixed ANOVA, with the former factor varying between-participants and the latter varying within-participants. Both main effects were significant, as was the hypothesized interaction,  $F(2,175) = 617.26, p < .001, \eta_p^2 = .876$ . Simple effects tests showed that explicit liking of Francis did not differ between story conditions at time 1,  $F(1,176) = .127, p = .722$ , but

that Francis was significantly more liked at time 2 in the fire rescue condition ( $M = 6.18$ ,  $SD = .98$ ) than in the control condition ( $M = 1.11$ ,  $SD = .27$ ),  $F(1,176) = 2181.37$ ,  $p < .001$ ,  $\eta_p^2 = .925$ . At time 3, Francis was also more liked in the fire rescue ( $M = 6.00$ ,  $SD = 1.21$ ) than the control condition ( $M = 1.43$ ,  $SD = .99$ ),  $F(1,176) = 760.82$ ,  $p < .001$ ,  $\eta_p^2 = .812$ . Figure 7 illustrates the explicit liking of Francis West at each measurement time in both story conditions.

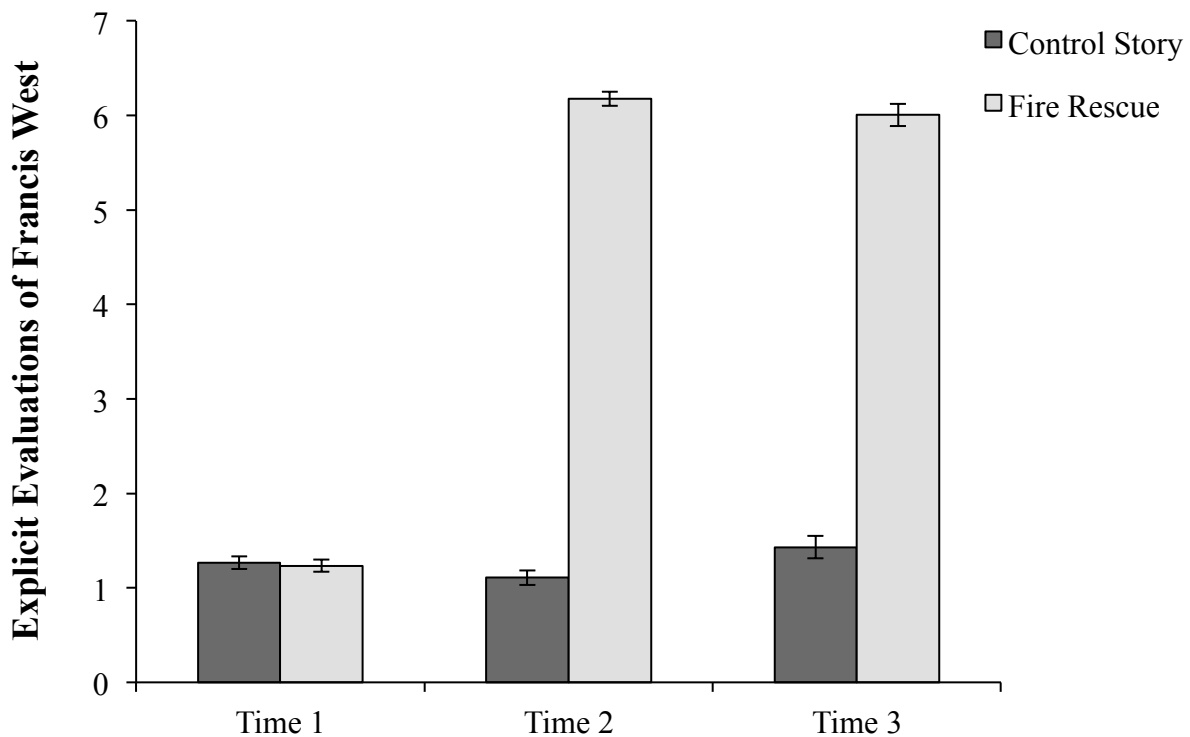


Figure 7. Mean explicit liking of Francis West in each story condition at each measurement time in Experiment 6. Error bars are standard errors.

## Discussion

After three days, implicit evaluations were still significant and positive (and thus revised from time 1) in the fire rescue condition, and still significant and negative (i.e., unchanged) in the control condition. Participants were not even re-exposed to his image until it was primed during the final AMP, and his name was not a prime on the task. This durability of the implicit

evaluations supports our claim that these revised evaluations are not merely transient effects of the testing situation that will revert back to their initial levels after the passing of a bit of time.

### General Discussion

Across seven studies, we found strong and consistent evidence that initial implicit evaluations can be undone. Importantly, we identified one factor capable of prompting such revision: reinterpretation of all the knowledge that formed the basis of the initial evaluation. In Experiments 1a and 1b, participants who had formed an initial negative implicit evaluation toward Francis West fully reversed those evaluations when they learned a reason for his bad behavior.<sup>12</sup> Experiment 2 showed that this reversal occurred only when participants read that his prior actions were aimed at saving children from a fire, but not when they read about a similarly positive action that did not explain his prior behavior (saving a baby from an oncoming train). This suggests that this revision depends on a recognition of the relationship between this new information and the old. Experiment 3 showed that this recognition requires at least minimal cognitive resources by demonstrating that high cognitive load at the time of reading the new information, relative to low load or no load, prevented full revision. Experiment 4 showed that subjective change in meaning of the earlier details of the story predicted more positive final implicit evaluations of Francis West in the fire rescue condition, and also showed that change in

---

<sup>12</sup> As one anonymous reviewer pointed out, in most of our experiments in the fire rescue condition there was not only a significant increase in the proportion of “pleasant” responses on trials with Francis West primes from time 1 to time 2, but a significant *decrease* in that proportion on trials with neutral face primes. We see two possibilities for why this occurs. One is that this may be simply an artifact: only one prime is highly relevant (Francis West), and the others are distractors of a sort. This may amplify the tendency to respond on those trials in an opposite way compared with the Francis West trials, and would be especially pronounced if a participant is trying to use the two response keys relatively evenly (see Scherer & Lambert, 2009, for an investigation of such contrast effects in priming tasks). Another possibility is that this effect demonstrates revision due to implicit social comparison. We view this as an exciting theoretical possibility. But, regardless, this effect does not pertain to our main finding of interest, which is the implicit positivity or negativity of Francis West relative to the control trials.

this paradigm is specific to the degree to which participants reinterpreted the earlier story rather than how much they thought about the information more generally. Experiment 5 demonstrated that reported change in the meaning of the earlier story details mediated the enhanced revision in the fire condition relative to the subway condition, which presented equally positive, but non-reinterpreting, information. Once again, a more general measure of elaborative thinking was not a significant mediator. Finally, Experiment 6 showed that the implicit evaluations formed in both the control and fire rescue conditions were not fleeting. After three days, implicit positive and negative evaluations were still apparent in the fire rescue and control conditions, respectively.

Collectively, these studies represent a closer examination of the conditions under which durable revision of implicit evaluations is possible, and identify one mechanism: reinterpretation. The empirical record to date has suggested that implicit evaluations are often resistant to efforts to undo a prior impression, even mere minutes after their initial formation (Gregg et al., 2006; Wilson et al., 2000). When they have shifted at all, it has been toward neutrality or ambivalence (Boucher & Rydell, 2012; Petty et al., 2006, Study 1; Peters & Gawronski, 2011, Experiment 3), or not occurred unless participants were compelled to engage in substantial elaboration (Wyer, 2010). Recent work by Cone and Ferguson (in press) found that initial positive implicit evaluations can indeed be overturned by new and highly diagnostic negative information, but the precise cognitive mechanism through which such change occurs, as well as whether revision could overturn an initial negative implicit evaluation, remained to be examined. Our work both shows negative-to-positive change in implicit evaluations and identifies a powerful mechanism driving it: not only must new information imply the opposite evaluation of the target, but the initial information must also be reframed. In other words, the reinterpretation that we used both invalidated the initial learning, and replaced it with new meaning.

In combination, our studies suggest that reversal of implicit evaluations can occur through reinterpretation, we have characterized some of the features of this process, finding it to be deliberate enough to require at least minimal cognitive resources but not to be interchangeable with just *any* extensive thinking about the story information. Although the task of identifying the complete set of requirements for implicit evaluation change through reinterpretation will extend beyond this set of studies, we now turn to a discussion of theoretical implications.

### **Reconciliation: What was Different This Time?**

The present studies identify the role of reinterpretation in implicit evaluation revision. However, one might note that in prior studies that failed to find full reversal, researchers similarly attempted to make appeals to the irrelevance of the initial information (e.g., Gregg et al., 2006; Peters & Gawronski, 2011, Experiment 3; Petty et al., 2006, Study 2). There are numerous possibilities for why our paradigm showed reversal while prior similar attempts did not. For one, the instructions in these prior studies prompted a reinterpretation of the old information that was somewhat different from our version of reinterpretation. Whereas in our studies the actions of Francis are no longer negative in light of the revelations about the situation, in those prior studies they suggested that the *targets* of the old information should be changed. In other words, in prior studies, the actions of aggressors (for example) are still negative, but just do not correspond to the group or person one initially thought they did. This type of reinterpretation may not be as easily implemented as in our case where an understanding of the new information basically compels the overturning of the initial impression (i.e., If Francis West was trying to save those kids, then his earlier actions were highly likely to have been enacted for that effort). It also may be that trying to realign behavior with new targets is ineffective at eliminating all traces of the initial information, as is the case with directed forgetting (Bjork & Bjork, 2003) or



adding new information to the old but not replacing it, as in the formation of contextualized evaluations (Gawronski & Cesario, 2013) and implicit ambivalence (Petty et al., 2006). Lastly, studies in which the targets change over time might also not produce as unified an impression as the paradigm we used here (e.g., Rydell et al., 2007). Ultimately, however, the task of identifying the differences between our paradigm and previous paradigms remains to be taken up.

Our findings offer new empirical support for the theoretical claim that reason-based routes to implicit evaluation revision should be possible. For instance, under the APE model (Gawronski & Bodenhausen, 2006, 2011), propositional reasoning is assumed to be capable of changing associative structure when new information is validated, but the parameters of this route of change have not been fully specified. Some studies have assumed that this change operates through the associative pairing of information contained within the propositions (Peters & Gawronski, 2011). But, the APE model does not outline when some kinds of affirmations will be more effective than others. Our results suggest that the effect of affirmation of new information will be especially strong when the new information *recasts* prior details, and that this process can produce full revision to the point where little evidence of the prior evaluation remains. Indeed, this recasting is able to force revision even in the case of an initial negative implicit evaluation turning to positive, an effect that has been particularly difficult to obtain (Cone & Ferguson, in press).

Likewise, under the metacognitive model (Petty et al., 2007), new “false” tags on old associations are assumed to be able to negate beliefs, but the conditions under which this occurs remain unspecified. Although the results in Wyer (2010) suggested that revisiting the prior details in light of the new information was necessary for the initial evaluation to be undone, the mechanism of change was unclear. Our results suggest that reinterpretation can lead to revision

without re-presenting the initial information, and pinpointed reinterpretation in particular as the type of reasoning that drives change in this paradigm, beyond more general elaborative thinking. As such, our results can be read as expanding the routes of revision of implicit evaluations under current theories, providing evidence of when and how such change occurs.

### **Implications for Established Evaluations**

Despite our results, a consistent finding in research on implicit evaluations is that people's implicit evaluations can be at odds with what they explicitly believe (Banaji & Greenwald, 2013). Our results do not imply that any and all implicit evaluations can be easily and rapidly changed through reasoning (reinterpretation or otherwise), just that there may be routes through which established implicit evaluations can be changed which have not been highlighted before, and which future investigations may profitably explore. Our view is that although we have explored implicit evaluation change in a specific scenario, we suspect that the mechanisms it illuminates operate in a variety of more mundane settings. That is, any time that a person *construes* new information to require a reinterpretation of prior knowledge about an evaluation object, we would expect shifts to occur in implicit evaluations proportionate to the amount and extremity of reinterpretation. An important future direction for this line of work is to examine situations in which this will be true. We examine some of these considerations below.

**Insight into the basis of the initial implicit evaluation.** In our studies, it is safe to assume that participants are quite aware of the basis for their evaluative feelings toward the target person, Francis West. However, with evaluations formed over a long period of time, people may have relatively poor introspective access to which content in memory actually affects their implicit evaluations. Insight into which information shapes judgments and impressions is often poor and based on inaccurate inferences or selective sampling of reasons (Wilson, Dunn,

Kraft, & Lisle, 1989). Although the claim that implicit evaluations are inaccessible to consciousness (e.g., Greenwald & Banaji, 1995) has been challenged (Gawronski, Hofmann, & Wilbur, 2006; Gawronski, LeBel, & Peters, 2007; Hahn et al., 2013), it remains likely that similar to the lack of source awareness of explicit evaluations (Bornstein, 1989; Hovland, Lumsdaine, & Sheffield, 1949; Kumkale & Albarracín, 2004; Wilson et al., 1989; Zajonc, 1968), people are not always aware of or may not remember the sources of their implicit evaluations (e.g., Dijksterhuis, 2004; Olson & Fazio, 2001, 2002; Rydell et al., 2006; see Gawronski et al., 2006). Thus, the reasons for an evaluation that people bring to mind when reflecting on their impressions of a person or group may differ from those that actually guide behavior. If rejection or reinterpretation of the basis of an initial evaluation is capable of revising even established implicit evaluations, as we posit here, this may be difficult if those sources in memory are forgotten, were never known, or are inaccurately identified (see also Lane, Ryan, Nadel, & Greenberg, in press). To the extent that they are able to identify reasons that *do* contribute to their current implicit evaluation, they may be able to enact a greater amount of revision than they could otherwise, a possibility for future investigation.

**Lack of information that prompts a full reinterpretation.** In the Francis West paradigm, the scenario is designed to allow for a single piece of new information to completely recast everything that was previously learned about the target person. One explanation for why implicit evaluation change might be more difficult for many established evaluations is that this type of new information may not be encountered. But, such cases exist, such as when a Nazi seems like he is supporting the holocaust when in fact he is saving over 1,000 people. In fact, any situation in which ulterior motives come to light is a potential case for revision of first impressions to occur.

**Motivational considerations.** Another relevant consideration is that people are not always willing or able to process information in an unbiased way (Frey, 1986; Kunda, 1990; Lord, Ross, & Lepper, 1979). Under such circumstances, even when information that contradicts a current position is attended, it is often held to a higher bar than information that supports one's position (Ditto & Lopez, 1992; Ditto et al., 2003; Eagly, et al., 2000).

The Francis West paradigm all but compels participants to change their minds about the target person; preserving the initial evaluation is quite indefensible. Though this may sometimes be the case in real life (e.g. discovering someone to be the victim of false and malicious accusations), it may be more typically the case that the effect of new information on an existing impression is a function less of the properties of that information itself than the manner and extent to which it is elaborated (Greenwald, 1968; Petty & Cacioppo, 1986; Petty, Ostrom, & Brock, 1981). Often, sweeping recalibration of past beliefs about an evaluation object may primarily occur when one is motivated to construe new information as prompting such revision, rather than new information inherently *requiring* such recalibration to occur.

### **Implicit vs. explicit evaluations: How do they relate to one another?**

The central theoretical contribution of the present work is the demonstration that changes to the meaning of prior information can lead to a full reversal of previously learned implicit evaluations. And yet, what does this mean for the presumed relations between implicit and explicit evaluations? After all, claims about (two) different processes underlying each type of evaluation have been used to explain the many examples of dissociation among them (e.g., Gawronski & LeBel, 2008; Gawronski & Strack, 2004; Petty et al., 2006; Rydell et al., 2006). If the processes underlying them are not as distinct as assumed – as our findings might imply – then why do we see so much evidence for dissociation elsewhere? One potential explanation for

such dissociations concerns the lack of “structural fit” among explicit versus implicit measures, including features such as format, stimuli, instructions, etc. (Payne, Burkley, & Stokes, 2008). These differences could explain dissociation without necessarily invoking any claims about underlying processes. Although there is a small amount of data suggesting that implicit and explicit evaluations differ even when controlling for fit (see Payne et al., 2008), this remains an open empirical question.

Another consideration is the tendency to assume that differences in behavior are due to dissociated processes. It is difficult (for us) to think of any behavioral evidence that would alone adjudicate between propositional versus associative processing, because any behavioral evidence can always be explained by boot-strapped versions of one’s favorite propositional or associative account (see Ferguson, Mann, & Wojnowicz, 2014; Moors, 2014). What we can do, however, is specify the circumstances under which implicit and explicit evaluations form, change, and predict behavior. One can then create computational models that formally test theories of associative versus propositional processing, as has been done frequently in cognitive psychology (e.g., Botvinick & Plaut, 2006; Read & Montoya, 1999; Sun, Slusarz, & Terry, 2005). For now, we have demonstrated one way in which implicit evaluations can be completely undone through a propositional, or reason-based, route. What remains is the work of figuring out what these findings mean for how implicit and explicit evaluations relate to one another.

## **Conclusion**

Implicit evaluations are not immune to revision through reason. Far from being “stuck” in dogged opposition to our reasoned conclusions about the validity of prior impressions, our implicit evaluations can reflect our updated interpretations of the world. Our findings suggest that to change unwanted implicit evaluations, we may marshal reason to undermine the *bases* of

our evaluations, if we can identify and edit them. Future work can examine other routes to implicit evaluation change and identify the conditions under which they are successful.

## References

- Amodio, D. M. (2014). Dual experiences, multiple processes: Looking beyond dualities for mechanisms of the mind. In J. S. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual process theories of the social mind* (pp. 560-576). New York: Guilford Press.
- Amodio, D. M., & Ratner, K. G. (2011). A memory systems model of implicit social cognition. *Current Directions in Psychological Science*, 20(3), 143–148.
- Banaji, M. R. & Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. Random House.
- Barden, J., & Tormala, Z. L. (2014). Elaboration and Attitude Strength: The new meta-cognitive perspective. *Social and Personality Psychology Compass*, 8(1), 17–29.
- Bar-Anan, Y., & Nosek, B. A. (2012). Reporting intentional rating of the primes predicts priming effects in the affective misattribution procedure. *Personality and Social Psychology Bulletin*, 38, 1193-1207.
- Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5, 323–370.
- Bjork, E. L., & Bjork, R. A. (2003). Intentional Forgetting can increase, not decrease, the residual influences of to-be-forgotten information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 524-531.
- Blair, I. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychological Review*, 6, 242-261.
- Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, 81, 828 – 841.

- Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin*, 106, 265–289.
- Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review*, 113(2), 201–233.
- Boucher, K. L., & Rydell, R. J. (2012). Impact of negation salience and cognitive resources on negation during evaluation formation. *Personality and Social Psychology Bulletin*, 38, 1329-1342.
- Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian conditioning. *Psychological Bulletin*, 114, 80–99.
- Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning & Memory*, 11(5), 485–494.
- Bouton, M. E., Westbrook, R. F., Corcoran, K. A., & Maren, S. (2006). Contextual and temporal modulation of extinction: Behavioral and biological mechanisms. *Biological Psychiatry*, 60(4), 352–360.
- Briñol, P., Petty, R. E., & Mccaslin, M. J. (2009). Changing evaluations on implicit versus explicit measures: What is the difference? In R. E. Petty, R. H. Fazio, & P. Brinol (Eds.), *Insights from the new implicit measures* (pp. 285-326). New York: Psychology Press.
- Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond bipolar conceptualizations and measures: The case of evaluations and evaluative space. *Personality and Social Psychology Review*, 1, 3–25.
- Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of



- implicit social cognition: A meta-analysis of associations with behavior and explicit evaluations. *Personality and Social Psychology Review*, 4, 330-350.
- Cone, J., & Ferguson, M. J. (in press). He Did *What*? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, doi: 10.1037/pspa0000014.
- Conrey, F. R., & Smith, E. R. (2007). Evaluation representation: Evaluations as patterns in a distributed, connectionist representational system. *Social Cognition*, 25(5), 718–735.
- Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology*, 89(4), 469–487.
- Crowe, D. M. (2004). *Oskar Schindler: The untold account of his life, wartime activities, and the true story behind the list*. Cambridge, MA: Westview Press.
- Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic evaluations: Combating automatic prejudice with images of liked and disliked individuals. *Journal of Personality and Social Psychology*, 81, 800-814.
- De Houwer, J. (2006). Using the Implicit Association Test does not rule out an impact of conscious propositional knowledge on evaluative conditioning. *Learning and Motivation*, 37(2), 176–187.
- De Houwer, J. D. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, 8(7), 342–353.
- De Houwer, J., Gawronski, B., & Barnes-Holmes, D. (2013). A functional-cognitive framework for evaluation research. *European Review of Social Psychology*.
- Deutsch, R., Gawronski, B., & Strack, F. (2006). At the boundaries of automaticity: Negation as

- reflective operation. *Journal of Personality and Social Psychology*, 91, 385-405
- Dijksterhuis, A. (2004). I like myself but I don't know why: Enhancing implicit self-esteem by subliminal evaluative conditioning. *Journal of Personality and Social Psychology*, 86, 345–355.
- Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, 63, 568–584.
- Ditto, P. H., Munro, G. D., Apanovitch, A. M., Scepansky, J. A., & Lockhart, L. K. (2003). Spontaneous skepticism: The interplay of motivation and expectation in responses to favorable and unfavorable medical diagnoses. *Personality and Social Psychology Bulletin*, 29, 1120–1132.
- Eagly, A. H., Kulesa, P., Brannon, L. A., Shaw, K., & Hutson-Comeaux, S. (2000). Why counterattitudinal messages are as memorable as proattitudinal messages: The importance of active defense against attack. *Personality and Social Psychology Bulletin*, 26(11), 1392–1408.
- Fazio, R. H. (2007). Evaluations as object-evaluation associations of varying strength. *Social Cognition*, 25, 603–637.
- Ferguson, M. J., & Fukukura, J. (2012). Likes and dislikes: A social cognitive perspective. In S. Fiske, & C. N. Macrae (Eds.), *Sage Handbook of Social Cognition* (pp. 165-189). Los Angeles: SAGE.
- Ferguson, M. J., Mann, T. C., & Wojnowicz, M. (2014). Rethinking duality: Criticisms and ways forward. In J. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual process theories of the social mind* (pp. 578-594). Guilford Press.

- Ferguson, M. J., & Wojnowicz, M. T. (2011). The when and how of evaluative readiness: A social cognitive neuroscience perspective. *Social and Personality Psychology Compass*, 5(12), 1018–1038.
- Frey, D. (1986). Recent research on selective exposure to information. *Advances in Experimental Social Psychology*, 19, 41–80.
- Gast, A., & De Houwer, J. (2013). The influence of extinction and counterconditioning instructions on evaluative conditioning effects. *Learning and Motivation*, 44(4), 312–325.
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit evaluation change. *Psychological Bulletin*, 132, 692–731.
- Gawronski, B., & Bodenhausen, G. V. (2011). The associative-propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology*, 44, 59–127.
- Gawronski, B., & Cesario, J. (2013). Of mice and men: What animal research can tell us about context effects on automatic responses in humans. *Personality and Social Psychology Review*, 17(2), 187–215.
- Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When “just say no” is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology*, 44, 370–377.
- Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are “implicit” evaluations unconscious? *Consciousness and Cognition*, 15, 485–499.
- Gawronski, B., & LeBel, E. P. (2008). Understanding patterns of evaluation change: When

- implicit measures show change, but explicit measures do not. *Journal of Experimental Social Psychology*, 44, 1355-1361.
- Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us?: Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science*, 2(2), 181-193.
- Gawronski, B., Rydell, R. J., Vervliet, B., & De Houwer, J. (2010). Generalization versus contextualization in automatic evaluation. *Journal of Experimental Psychology: General*, 139, 683-701.
- Gawronski, B., & Strack, F. (2004). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology*, 40(4), 535-542.
- Gilbert, D. T., & Osborne, R. E. (1989). Thinking backward: Some curable and incurable consequences of cognitive busyness. *Journal of Personality and Social Psychology*, 57(6), 940-949.
- Greenwald, A. G. (1968). Cognitive learning, cognitive response to persuasion, and evaluation change. In A. G. Greenwald, T. C. Brock, and T. M. Ostrom (Eds.), *Psychological foundations of evaluations* (pp. 147-170). New York: Academic Press.
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Evaluations, self-esteem, and stereotypes. *Psychological Review*, 102, 4-27.
- Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (in press). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*.
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual

- differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464–1480.
- Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, 85(2), 197–216.
- Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97(1), 17–41.
- Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, 90(1), 1–20.
- Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2013). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, 143(3), 1369–1392.
- Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis*. New York: Guilford Press.
- Horcajo, J., Briñol, P., & Petty, R. E. (2010). Consumer persuasion: Indirect change and implicit balance. *Psychology and Marketing*, 27(10), 938–963.
- Hovland, C. I., Lumsdaine, A., & Sheffield, F. (1949). *Experiments on mass communication*. Princeton, NJ: Princeton University Press.
- Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). The dominance of associative theorizing in implicit attitude research: Propositional and behavioral alternatives. *The Psychological Record*, 61(3), 6.
- Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of*

- Personality and Social Psychology*, 81(5), 774–788.
- Knobe, J. (2006). The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies*, 130, 203–31.
- Kumkale, G. T., & Albarracín, D. (2004). The sleeper effect in persuasion: A meta-analytic review. *Psychological Bulletin*, 130, 143–172.
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108, 480–498.
- Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. L., Joy-Gaba, J. A., ... & Nosek, B. A. (in press). A comparative investigation of 17 interventions to reduce implicit racial preferences. *Journal of Experimental Psychology: General*.
- Lane, R. D., Ryan, L., Nadel, L., & Greenberg, L. (in press). Memory reconsolidation, emotional arousal and the process of change in psychotherapy. *Behavioral and Brain Sciences*.
- Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and evaluation polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11), 2098–2109.
- Malle, B. F., & Knobe, J. (1997). Which behaviors do people explain? A basic actor-observer asymmetry. *Journal of Personality and Social Psychology*, 72, 288–304.
- McNulty, J. K., Olson, M. A., Meltzer, A. L., & Shaffer, M. J. (2013). Though they may be unaware, newlyweds implicitly know whether their marriage will be satisfying. *Science*, 1119–1120.
- Moors, A. (2014). Examining the mapping problem in dual process models. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pg. 20–34). Guilford Press.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on

- mental processes. *Psychological Review*, 84, 231-259.
- Olson, M. A., & Fazio, R. H. (2001). Implicit evaluation formation through classical conditioning. *Psychological Science*, 12, 413–417.
- Olson, M. A., & Fazio, R. H. (2002). Implicit acquisition and manifestation of classically conditioned evaluations. *Social Cognition*, 20, 89–104.
- Olson, M. A., & Fazio, R. H. (2006). Reducing automatically-activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin*, 32, 421-433.
- Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, 105(2), 171–192.
- Payne, B. K., Brown-Iannuzzi, J., Burkley, M., Arbuckle, N. L., Cooley, E., Cameron, C. D., & Lundberg, K. B. (2013). Intention invention and the affect misattribution procedure: Reply to Bar-Anan and Nosek (2012). *Personality and Social Psychology Bulletin*, 39(3), 375–386.
- Payne, B. K., Burkley, M., & Stokes, M. B. (2008). Why do implicit and explicit evaluation tests diverge? The role of structural fit. *Journal of Personality and Social Psychology*, 94, 16-31.
- Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. (2005). An inkblot for evaluations: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89, 277-293.
- Perugini, M., Richetin, J., & Zogmaister, C. (2010). Prediction of behavior. In B. Gawronski &

- B. K. Payne (Eds), *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications* (pp. 256-278). New York: Guilford Press.
- Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin*, 37, 557-569.
- Petty, R. E. & Briñol, P. (2010). Evaluation structure and change: Implications for implicit measures. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp.335-352). New York: Guilford Press.
- Petty, R. E., Briñol, P., & DeMarree, K. G. (2007). The meta-cognitive model (MCM) of evaluations: Implications for evaluation measurement, change, and strength. *Social Cognition*, 25, 657-686.
- Petty, R. E., & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123-205). New York: Academic Press.
- Petty, R. E., Haugtvedt, C. P., & Smith, S. M. (1995). Elaboration as a determinant of attitude strength: Creating attitudes that are persistent, resistant, and predictive of behavior. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 93–130). Mahwah, NJ: Erlbaum.
- Petty, R.E., Ostrom, T.M., & Brock, T.C. (1981). Historical foundations of the cognitive response approach to evaluations and persuasion. In R. Petty, T. Ostrom, & T. Brock (Eds.), *Cognitive responses in persuasion* (pp. 5-29). Hillsdale, NJ: Erlbaum.
- Petty, R. E., Tormala, Z. L., Briñol, P., & Jarvis, W. B. G. (2006). Implicit ambivalence from



- evaluation change: An exploration of the PAST model. *Journal of Personality and Social Psychology*, 90(1), 21–41.
- Read, S. J., & Montoya, J. A. (1999). An autoassociative model of causal reasoning and causal learning: Reply to Van Overwalle's (1998) critique of Read and Marcus-Newhall (1993). *Journal of Personality and Social Psychology*, 76(5), 728-742.
- Reeder, G. D., Pryor, J. B., & Wojciszke, B. (1992). Trait-behavior relations in social information processing. In G. R. Semin, & K. Fielder (Eds.), *Language, interaction and social cognition* (pp. 37-57). Thousand Oaks, CA: Sage Publications.
- Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5, 296 –320.
- Rule, N. O., Tskhay, K. O., Freeman, J. B., & Ambady, N. (2014). On the interactive influence of facial appearance and explicit knowledge in social categorization. *European Journal of Social Psychology*.
- Rydell, R. J., & Gawronski, B. (2009). I like you, I like you not: Understanding the formation of context-dependent automatic evaluations. *Cognition and Emotion*, 23, 1118-1152.
- Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit evaluation change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, 91(6), 995–1008.
- Rydell, R. J., McConnell, A. R., Mackie, D. M., & Strain, L. M. (2006). Of two minds: Forming and changing valence-inconsistent implicit and explicit evaluations. *Psychological Science*, 17(11), 954–958.
- Rydell, R. J., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2007).

- Implicit and explicit evaluations respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology*, 37(5), 867–878.
- Scherer, L. D., & Lambert, A. J. (2009). Contrast effects in priming paradigms: Implications for theory and research on implicit attitudes. *Journal of Personality and Social Psychology*, 97(3), 383-403.
- Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & Groom, C. J. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review*, 115(2), 314–335.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (October 14, 2012). A 21 word solution. Available at SSRN: <http://ssrn.com/abstract=2160588>
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119(1), 3-22.
- Smith, C. T., De Houwer, J., & Nosek, B. A. (2013). Consider the source: Persuasion of implicit evaluations is moderated by source credibility. *Personality and Social Psychology Bulletin*, 39(2), 193–205.
- Steinhouse, H. (1994). “The real Oskar Schindler.” *Saturday Night*, 109(3), 40-45+. Retrieved from <http://www.writing.upenn.edu/~afilreis/Holocaust/steinhouse.html>
- Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review*, 112(1), 159–192.
- Towles-Schwen, T., & Fazio, R. H. (2006). Automatically-activated racial attitudes as predictors of the success of interracial roommate relationships. *Journal of Experimental Social Psychology*, 42, 698-705.
- Wegner, D. M. (1994). Ironic processes of mental control. *Psychological Review*, 101(1), 34-52.

- Wilson, T. D., Dunn, D. S., Kraft, D., & Lisle, D. J. (1989). Introspection, evaluation change, and evaluation-behavior consistency: The disruptive effects of explaining why we feel the way we do. In L. Berkowitz (Ed.). *Advances in experimental social psychology* (Vol. 22, pp. 287–343). New York: Academic Press.
- Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual evaluations. *Psychological Review*, 107(1), 101-126.
- Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: Variability in automatically activated evaluations. *Journal of Personality and Social Psychology*, 81, 815–827.
- Wyer, N. A. (2010). You never get a second chance to make a first (implicit) impression: The role of elaboration in the formation and revision of implicit impressions. *Social Cognition*, 28(1), 1–19.
- Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9, 1–27.

## Appendix A: Story Details Across Studies

### Time 1 Story Sentences

- 1) **[Experiment 1b only]** Francis' small town was about 99% white, but recently the Griffins and Wards, the town's first ever interracial families, had moved in.
- 2) Francis West knew that the couples that owned the two other homes on his street, the Griffins and Wards, had gone out for a bit.
- 3) After a few minutes, he looked out the window at the Griffin house, and decided that he just had to go over there.
- 4) Francis grabbed an axe from his cellar and went across the backyard to the Griffins' porch door.
- 5) He started hacking away at the door, cursing and occasionally kicking, as the door groaned and finally yielded under the force of Francis' blows.
- 6) The door splintered into a million pieces, ruining a tiny painting of a butterfly that the Griffins' daughter Zoe had painted on its lower left corner only days before.
- 7) Treading mud onto the Griffins' freshly installed beige carpet, Francis made his way through the family room of the home.
- 8) As he went, Francis knocked over and shattered a set of priceless vases handed down to Mrs. Griffin by her grandmother. The pieces scattered on the floor, mixing with the mud from Francis' boots.
- 9) As he entered the kitchen, Francis spotted a large pot full of water sitting on top of the stove. He grabbed the pot and proceeded to throw the water all over the kitchen, drenching and destroying Zoe's first laptop, which was sitting on the countertop.

- 10) Francis took the pot with him as he moved on into the stairwell to the second floor, throwing the remaining water all over the hallway. Much of it doused a painting that Mrs. Griffin's mother had made for her years before her passing; it was Mrs. Griffin's favorite.
- 11) As Francis arrived at the second floor, he methodically searched the bedrooms for precious things, stomping all over some pictures that young Zoe had left on the floor by the top of the stairs. Francis didn't even care.
- 12) Francis found what he was looking for, and left the house with it.
- 13) After leaving the Griffin house, Francis identified the adjoining Ward home as his next target.
- 14) Traipsing right through Mrs. Ward's prized garden, destroying years worth **[Experiment 1b: countless hours]** of careful cultivation, Francis arrived at the big bay window next to the back door.
- 15) The window was adorned with home-crafted stained glass; Francis thought nothing of smashing right through it with a brick from a nearby pile.
- 16) As he climbed through the Wards' window, Francis knocked over the family's large big-screen TV, which crashed roughly onto the hardwood floor.
- 17) As he moved from the family room to the hallway, Francis stepped on the Wards' cat Paws, which squealed and fled off down the hall.
- 18) Francis awkwardly made his way up the stairs, often swaying from side to side, and knocked down several pieces of ceramic art that the Wards had on display alongside their staircase. And Francis didn't care.
- 19) As he arrived on the second floor, Francis made his way through the bedrooms, as he did at the Griffins, looking for precious things.

- 20) Finding what he was looking for, Francis turned back to the stairs.
- 21) Looking toward the front door, Francis saw the cat Paws lying dead in front of it. He didn't really care.
- 22) Francis slowly made his way toward the basement door, and after descending the stairs, started kicking and shoving boxes and things left and right, like a madman.
- 23) The prized china that filled a few boxes shattered under the force of Francis' kicks.
- 24) Francis stepped on and walked across an open bin of family photos, spreading mud all over some of the baby pictures of the family's young son Mark.
- 25) Reaching the basement door, Francis roughly shoved it open and moved out into the yard. A few family photos stuck to the heel of his shoes.
- 26) Francis moved toward the sidewalk.
- 27) He faced the two houses, now quite damaged, and sat there with the things he had taken from them. He looked down the road and waited for the return of the Griffins and Wards.

### **Time 2 Story Sentences**

#### **Fire Rescue Condition [All experiments]:**

Francis West broke into the adjoining Griffin and Ward homes because he saw that they were on fire. The only precious things he removed from either home were the young kids Zoe and Mark, and he waited on the sidewalk with them until their worried parents' return.

#### **Control Condition [Experiments 1a, 1b, 2, and 3]:**

While he waited on the sidewalk to confront the Wards and Griffins, Francis started picking up rocks from the roadside and hurling them at the houses. By the time the horrified families returned, nearly all of the windows had been smashed by rocks, and dents covered the front of both houses.

**Control Condition [Experiment 6]:**

It turned out that Francis West had been arrested previously for multiple crimes, including armed robbery and physical assault. Neighbors reported that he often screamed at neighborhood kids, and had yelled at Zoe and Mark for playing tag near the corner of his property the previous day. He apparently trashed the families' homes in search of valuables, as well as in revenge.

**Subway Rescue Condition [Experiments 2 and 5]:**

At a different point in time, Francis West was in the news because he was at a subway station when he noticed that a baby had crawled and fallen onto the tracks below. Seeing a rapidly approaching train, Francis jumped down onto the tracks, grabbed the baby, and climbed up to safety a split-second before the train came roaring past.

### **Chapter III. Mechanisms and Boundary Conditions**

Although the work presented in Chapter II builds the case that initially negative implicit first impressions can be rapidly reversed, a critical next step is to ask, under what broader set of conditions can such revision occur? Given that efforts to attenuate implicit bias toward real stigmatized groups have been of limited success even in the short term (Lai et al., 2014) and even less effective after a delay (Lai et al., 2016; see also Peters & Gawronski, 2011), it is critical to identify the minimal and broadest conditions under which successful reversal of negative implicit first impressions can be achieved. The work in Chapter II did provide some evidence for boundaries and predictors of reinterpretation-driven reversal. First, I showed that reinterpretation information led to implicit revision only when participants were under low cognitive load (rehearsing two digits) or none. When they were under high load (rehearsing 8 digits), implicit revision was reduced. Second, participants' reported cognitive content predicted the degree of implicit revision. The more participants reported thinking about how the new information changed the meaning of the earlier events in the story, the more implicit revision they showed.

The evidence so far shows, then, that information that reinterprets the original information is a kind of evidence that produces robust implicit revision, and that it requires some moderate degree of cognitive resources. It remains unknown, however, how generalizable the findings of Chapter II might be to a wider variety of contexts. The goal of Chapter III is to broaden the investigation to begin to address that unknown.

One question about the generalizability of the kind of revision shown in Chapter II is, just how unique is reinterpretation-type evidence is in facilitating the revision of implicit responses? Is reinterpretation one instance of a broader class of strategies through which implicit



impressions can be readily updated, or is reinterpretation relatively exceptional in its effectiveness? Studies 7 and 8 will probe this question by conceptualizing reinterpretation in terms of more basic constituent process elements that might be shared by a larger family of strategies for producing implicit revision. The studies reveal evidence that supports the idea that other cases that pair a negation of the validity of earlier information about a person with new, countervailing information can be similarly effective to reinterpretation, and impact implicit responses in a similar manner.

Study 9 moves to another form of external validity by testing whether reinterpretation can produce updating even after days have passed since the original learning, an important test of the wider efficacy of reinterpretation given that impressions of real people may not be as immediately corrected as they were in the original studies of Chapter II (or most other work on implicit impression formation and change; e.g., Gregg et al., 2006; Peters & Gawronski, 2011; Rydell & McConnell, 2006). This study presents a paper published with coauthor Melissa J. Ferguson in the *Journal of Experimental Social Psychology*, titled “Reversing implicit first impressions through reinterpretation after a two-day delay.”

Finally, it is not yet clear how generally effective reinterpretation can be as a route of changing implicit impressions under a broader set of contexts. Though cases of clear, extreme, uncontestable reinterpretation – along the lines of the Francis West story – undoubtedly occur, and might be likely to produce strong implicit revision in light of the findings in Chapter II, there are many other times in real life in which reinterpretation might be milder. A reinterpretation might contradict a broader stereotype, might change only some of one’s prior reasoning about a person, or might simply be less extreme; in such cases, can reinterpretation still be effective?

Studies 10-12 will test the possibility of reinterpretation-driven change in the contexts of a contentious issue, racial prejudice, and gender stereotyping.

### **Reinterpretation and Other Routes to Revision**

In Studies 7-8, I sought to establish the general nature of information that can lead to a rapid revision of negative implicit first impressions, abstracting away from the particularities of reinterpretation. First, can other strategies that, like reinterpretation, involve a combined rejection of an initial impression and its replacement with a new one be just as effective in revising implicit responses? Second, do such “negation + replacement” strategies—including, but not limited to, reinterpretation—impact implicit impressions in a similar way, by changing the evaluative information associated with that person, vs. merely changing response biases? If reinterpretation and other information that involves joint negation and replacement are similarly effective in producing revision, and impact the processes underlying implicit impressions in a similar manner, this would provide evidence that a broader class of strategies involving the combined negation and replacement of initial impressions might be effective for updating implicit impressions. Below, I discuss my approach in more detail.

### **Reinterpretation as a Form of Negation + Affirmation**

Can any features of reinterpretation be abstracted in such a way as to suggest other strategies that might be similarly effective in reversing implicit evaluations? Possibly, reinterpretation may consist of at least two generalizable cognitive steps: first, the *negation* of an aspect of the initial information, and second, the *affirmation* (or, addition) of new information that supports an impression opposite in valence from the original one. With reinterpretation, the meaning of the initial information (and the first impression that it implies) is negated, and new details that support a different meaning (and an opposite impression) are affirmed. Past research

has examined these steps of negation and affirmation mostly in isolation, as reviewed in Chapter II under the terminology of “subtraction” and “addition”, and again here briefly.

**Negation.** Although negation – a rejection of the truth-value of previously learned information– can readily update *explicit* judgments, empirical findings and theoretical positions are more mixed about whether negations can reverse *implicit* impressions. On the one hand, some experiments have found that such negation attempts do not change implicit impressions (Gregg et al., 2006; see also Wilson et al., 2000), with some models (e.g., APE; Gawronski & Bodenhausen, 2006, 2011) arguing that the activation of associations that are posited to drive implicit impressions are unaffected by negations, which simply keep the associations that underlie the rejected information active (Gawronski, Deutsch, Mbirikou, Seibt, & Strack, 2008). On the other hand, some findings have suggested that under some circumstances, negation can impact implicit impressions (Boucher & Rydell, 2012; Deutsch, Kordts-Freudinger, Gawronski, & Strack, 2009; Johnson, Kopp, & Petty, in press; Peters & Gawronski, 2011). Peters and Gawronski (2011) found that when negations were presented partially simultaneously with the to-be-negated information during an impression formation task, they were effective in producing negation-consistent implicit impressions. However, a delay of even a few minutes made them less effective, suggesting that the window for effective negation may be restricted to the moment of learning. Boucher and Rydell (2012) found that negations can be effective if they are made particularly strong and salient at the time when they are processed, and suggested that a longer delay may lead to weaker integration of the negation. Consistent with the idea that the strength of negations can vary in impactful ways, recent work found that instructions to think “THAT’S WRONG!” in response to stereotype-consistent statements produced a reduction in implicit race-based prejudice, whereas instructions to think “NO!” did not (Johnson et al., in press). The

Metacognitive Model (MCM; Petty et al., 2006; Petty, Briñol, & DeMarree, 2007) accounts for such findings by proposing that negations can be directly reflected in the associations underlying implicit impressions, and other theories that assume that implicit evaluations are based in part or wholly on propositional representations (De Houwer, 2014; Hughes, Barnes-Holmes, & De Houwer, 2011) would also predict that negation should be able to change implicit evaluations. As this perspective holds that the representations underlying implicit evaluation are propositional, a negation stored in memory could directly impact implicit responses.

**Affirmation.** Another form of propositional reasoning that may drive implicit change is affirmation—judging information that implies a new impression to be true. Multiple studies have presented varying amounts of behavioral details that imply an opposite impression from the first impression (e.g., Gawronski et al., 2010; Petty et al., 2006; Rydell & Gawronski, 2009; Rydell & McConnell, 2006; Rydell et al., 2007), with mixed success. Studies have found that a large amount of relatively mild to moderate new information is needed to reverse implicit impressions (e.g., Rydell & McConnell, 2006). Similarly, Cone and Ferguson (2015) found that new, extreme information did succeed in revising implicit impressions, but *reversal* only occurred when going from positive first impressions to revised negative ones.

Such findings are consistent with models that argue that affirming a new proposition can lead to the activation of different associations; for example, if one currently has an active association between “Bob” and “bad” but then affirms the belief that “Bob is always kind,” this may activate different associations (“Bob-kind”) that might then drive implicit responding, associations which could then be strengthened (Gawronski & Bodenhausen, 2006, 2011). The MCM, on the other hand, holds that validity information can itself be represented as a “true” tag associated with an impression. As with the “false” tags resulting from negation, its strength will

be a product of extent and depth of elaboration (Petty et al., 2007; Wyer, 2010). Finally, propositional models (De Houwer, 2014; Hughes, Barnes-Holmes, & De Houwer, 2011) also predict that affirming new propositions reflecting countervailing information should lead to implicit revision of first impressions.

**Combining Negation + Affirmation.** The effect of reinterpretation on implicit evaluations might be effective at least in part because it draws upon these two cognitive processes of negation and affirmation. Either one alone may be inconsistently effective, as prior work has shown, but their combination may be the explanation for why reinterpretation seems so robust and effective.

The question is, do these two steps have to be linked, or related, in some way to be effective in driving implicit updating? It may be that reinterpretation works simply because the original evidence is negated, and then new countervailing evidence is affirmed. However, the alternative possibility is that the negation of the earlier information has to be *related* to the affirmation of the new information, as it is in reinterpretation types of evidence. That is, telling people that a man who broke into the homes of his neighbors was actually saving kids from a fire accomplishes these two steps in one fell swoop.

Is reinterpretation, then, relatively unique in its ability to revise negative implicit impressions, or is it a subtype of a broader set of effective routes of change that all involve the steps of negation and affirmation? Studies 7-8 were designed to test exactly this question. I presented participants with either negation or affirmation information alone, or a combination of negation and affirmation information. Moreover, I manipulated whether the negation was related or unrelated to the affirmation information. This design can then demonstrate, in one study, the necessary and sufficient conditions required for reversing impressions.

What do contemporary theories predict about the relative effectiveness of these types of evidence? Although they do not weigh in on whether the relatedness (of the negation and affirmation) matters, they do make claims about how negation + affirmation (related or not) should compare to the other conditions. For example, the APE model suggests that negation, when combined with affirmation, should not generally be more effective than affirmation alone—and may in fact be *less* effective than the latter. Under the APE model, the most straightforward explanation of the strong effects of reinterpretation on implicit impressions would likely be that the new details, which reframe earlier behavior in a new light, lead to the strong affirmation of a new proposition (e.g., “the man is a hero”), which builds a strong new association between the target and a positive evaluation. Though the reinterpretation could also be viewed as negating the earlier, erroneous, negative impression of the target’s actions, dwelling on this negation itself (“the man is *not* a criminal”) could not contribute to implicit change, because it would serve only to maintain activation of the association between the target and a negative impression. However, the negation element of reinterpretation is strongly implicated in revision, because the results of Chapter II found that equally positive new information that was unrelated to (and did not overturn) the earlier interpretation of the story did not similarly lead to a reversal of the implicit impression (consistent with evidence also from Cone & Ferguson, 2015).

On the other hand, the MCM would more readily accommodate the proposal that reinterpretation works through the joint operation of negation and affirmation: when participants notice the high relevance of the new information to the prior impression, they engage in thorough elaboration, if able, which makes the “false” tag produced by negation particularly strong (Wyer, 2010, 2016). The new impression will be linked with a strong “true” tag in the same way.

Finally, theories highlighting propositional reasoning (De Houwer, 2014; Hughes, Barnes-Holmes, & De Houwer, 2011) would also predict that negation plus affirmation would lead to implicit revision. Because this perspective proposes that propositional representations can directly drive implicit processes, it does arguably the best job of accommodating findings of fast revision through propositional negation and affirmation, as propositions of both of these types could be stored in memory and directly impact implicit responses once learned.

### **Comparing Strategies: Effects on Constituent Processes of Implicit Impressions**

Even if I find evidence in support of the idea that different types of information (beyond just reinterpretation) that prompt joint negation and affirmation can effectively update implicit impressions, this does not reveal anything about what precise effects the information might be having. Although it is clear that reinterpretation can alter the evaluative impressions captured by implicit measures (Chapter II), these responses are themselves the products of multiple processes, as no task is a process-pure measure of a single representation or process. The Implicit Association Test (IAT; Greenwald et al., 1998), for example, can be construed as tapping a combination of association activation, controlled detection of the target category, overcoming bias, and response bias (Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Sherman et al., 2008). Likewise, the Affect Misattribution Procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005)—which is the measure overwhelmingly used in the current work—can be understood as jointly tapping the activation of a response to a prime, the activation of a response to a target, and misattribution of the response triggered by the prime to the target (Payne, Hall, Cameron, & Bishara, 2010). Although the finding that reinterpretation and other forms of information that involve negation + replacement revise implicit impressions would be suggestive of a broader family of strategies for producing effective revision, it is possible for similar shifts

on an implicit measure to be produced in very different ways. On the AMP, shifts in implicit impressions can occur via changes in the evaluative impression activated by the face of the person (the “prime” in the AMP), but they can also be produced via shifts in other processes that are not as relevant, such as likelihood of misattributing, or the responses activated by the “target” in the AMP. If reinterpretation alters the impression activated by the prime stimulus but other negation + replacement information alters, say, the likelihood of misattribution, such evidence would strongly challenge the claim that reinterpretation can be abstracted into a broader mechanism of implicit revision through negation + replacement.

In the current work, I use multinomial process trees to model the effects of reinterpretation on the processes occurring during implicit evaluation (e.g., Batchelder & Riefer, 1999; Conrey et al., 2005; Jacoby, 1991; Payne, 2001; Payne et al., 2010; Sherman et al., 2008). Acknowledging that no behavioral response is process-pure (reflective of a single mental process), mathematical process models develop systems of equations designed to reflect the processes underlying the range of behavioral responses that are possible on a task, and then use the response data to estimate the degree to which these processes operate in different conditions. Mathematical process models have yielded numerous insights in social psychological research, such as that higher implicit bias in older adults is largely a function of diminished control rather than more negative associations to racial outgroups (Stewart, Von Hippel, & Radvansky, 2009), that differences in implicit pro-White/anti-Black evaluations between White and Black participants are driven primarily by differences in evaluative associations rather than controlled processing (Sherman et al., 2008), and that a public context both decreases controlled detection of correct responses and strengthens processes of overcoming bias (Conrey et al., 2005).



In Studies 7 and 8, I test whether the combined negation and replacement of an initial impression can lead to the updating of implicit impressions, and compare the effectiveness of that combination to reinterpretation. The overall hypothesis is that new information that prompts a combination of *negating* an earlier impression and *replacing* it with a new, countervailing one, will lead to the reversal of negative implicit first impressions, regardless of whether these two steps are intrinsically linked (in reinterpretation) or performed as separate steps. I test this possibility both by examining overall changes in implicit impressions in each condition (Study 7), and by using multinomial modeling to test for similarities and differences in the constituent processes that are affected (Study 8). I examine if both forms of negation + replacement lead to shifts in the alteration of evaluative responses that are unintentionally activated by a face, rather than the alteration of other processes that underlie the measurement of an implicit impression, such as misattribution. Note that prior to Study 7, I was agnostic on whether the reinterpretation and negate + affirm conditions would show equivalent revision, and remained open to the possibility that they would not; I thus view this study as relatively exploratory, and sought to replicate and extend its findings in Study 8.

### **Study 7: Components of Reinterpretation**

What are the “ingredients” of reinterpretation that make it effective in revising implicit impressions? Can reinterpretation be understood in terms of constituent mental processes—propositional negation and affirmation—such that information that prompts both steps *without* reinterpretation might be similarly effective? As reviewed above, reinterpretation seems to trigger a combination of “negating” an initial impression and simultaneously “replacing” it with another, in that the new behavioral details compel both. In the standard Francis West fire rescue condition, the new information not only suggests that the earlier impression of the target as a

villain is wrong, but also delivers an alternative – that he is a hero. An advantage to this view of reinterpretation is that it helps conceptualize these findings in closer proximity to current theories that draw on similar concepts (Gawronski & Bodenhausen, 2006, 2011; Petty et al., 2007).

In this study, I compared the effectiveness of information that prompts a reinterpretation to information that produces a joint negation and affirmation without reinterpretation. I made use of the Francis West materials first introduced in Chapter II, comparing the “fire rescue” reinterpretation condition to four other conditions: a control condition in which participants learned one more negative detail about the target (consistent with their prior impression), a “negate only” condition in which they were told that everything they had read previously about the target actually did not occur, an “affirm only” condition in which they were told that the target performed a heroic act unrelated to his earlier, negative behaviors, and a combined negate + affirm condition in which they were informed *both* that none of the earlier story had actually taken place and that instead, the target had performed an unrelated heroic action. To support the idea that combined negation and replacement strategies might be broadly effective in reversing implicit impressions whether they involve reinterpretation or not, it is critical that the negate + affirm information show greater revision than either the negate only information or affirm only information, and that the effect in this joint condition be similar to that in the reinterpretation condition.

## **Method**

**Participants and design.** I recruited 375 participants from Amazon’s Mechanical Turk (mturk.com) to participate in the study in return for \$1.75 ( $M_{\text{age}} = 32.31$  years,  $SD_{\text{age}} = 10.01$  years; 193 women, 181 men, 1 genderqueer). I planned a priori to recruit a total sample of this size to achieve approximately 75 participants per between-participants cell of the design, in the

range of prior studies and within budget constraints. An additional 8 participants began the study but did not complete it, so their partial results were not included in any analyses. Each participant was randomly assigned to one of 5 story conditions.

**Initial learning task.** After recruitment on Mturk, participants read through the Francis West story (Chapter II). Identical to the procedure in the earlier studies, for each participant, one image of the face of a white man was randomly selected from a set of 11 such images to be Francis West, and the remaining 10 were set aside to serve as control prime stimuli on the subsequent implicit measures. The images were drawn from a bank of face stimuli used in prior research (Minear & Park, 2004).

**First implicit measure.** After participants learned the information about Francis West, I assessed implicit evaluations of him using the AMP as presented in the studies within Chapter II. As before, 20 trials featured an image of Francis West as prime, and 20 trials featured one of the 10 control faces as prime. Comparing the proportion of “pleasant” responses between trials with different prime types thus allows for an assessment of implicit impressions of the primes (Payne & Lundberg, 2014; Payne et al., 2005, 2013; cf. Bar-Anan & Nosek, 2012).

**First explicit attitude scale.** After the AMP, participants completed the six-item scale assessing explicit evaluations of Francis West used throughout Chapter II.

**Second learning task.** Participants were next told that they would now read some additional information about Francis West, which depended on their story condition. In all conditions, participants were instructed to take at least 15 seconds to consider the new information, and then advanced to the next task at their own pace after at least that much time elapsed.

In the *control condition*, participants were informed that Francis performed one additional negative behavior in the context of the initial story. Specifically, they read that he began chucking rocks at the two houses that he had previously invaded.

In the *reinterpretation condition (fire rescue)*, participants read the information that reframed Francis' earlier actions from negative to positive (Chapter II). Specifically, they read that Francis had in fact entered the two homes because they were on fire, and the only "precious things" he removed from the bedrooms were the two young children who he knew were trapped inside. This new information was designed to produce a reinterpretation of the basis of their first impression of Francis, by reframing Francis' earlier actions from negative to positive.

In the *negate only condition*, participants were informed that everything they had previously read about Francis was actually false; he did not perform any of the actions described earlier.

In the *unrelated new information (affirm only; subway rescue)* condition, participants were told an additional detail about Francis West that described a positive behavior he had performed, but which was not directly related to the earlier information about Francis. Specifically, they read that Francis had once been at a subway station when a baby fell onto the tracks as a train approached; he heroically jumped down onto the tracks and climbed to safety with the baby mere moments before the train would have struck and killed them both.

In the *negate + affirm* condition, participants viewed a combination of the *negation* and *unrelated new information* descriptions – that everything they had previously read about Francis was false, and that instead, he had rescued a baby that fell onto the subway tracks as a train approached.

In all conditions, participants were instructed to take at least 15 seconds to consider the new information, and then advanced to the next task at their own pace after at least that much time elapsed.

**Second implicit measure.** Participants next completed the 40-trial AMP again, which was identical to the first except that the pictographs were drawn from the set not used in the first AMP.

**Second explicit attitude scale.** The same explicit attitude scale that participants previously completed was administered again.

**Final questionnaire items.** Finally, participants were asked whether they know Mandarin and/or Cantonese, to identify Francis West from the set of 11 face images presented during the study, to identify the final information they read about Francis West from short summaries of the five conditions, to provide their age and gender, and to give open-ended feedback regarding what they thought the study was about and any other comments they had. They were then debriefed, thanked, and compensated.

## Results

**Data preprocessing.** In accordance with established protocol for the AMP (Payne et al., 2005), I excluded *a priori* all participants who speak Mandarin and/or Cantonese, as the pictographs would not be neutral for such participants (2.7%), and all participants who used a single response key on every trial of at least one of the two AMPs, showing a disregard for instructions (5.9%). This left a final sample of 344 participants.

For both the Time 1 and Time 2 AMPs, for each participant I calculated the proportion of trials on which the pictograph was judged to be more pleasant than average following the Francis West prime image, and a similar proportion for trials following control face images, yielding

four AMP measures for each participant within a 2 (Time: time 1, time 2) x 2 (Prime: Francis West, Control) design. Explicit evaluations of Francis West at Time 1 and Time 2 were computed by averaging the six liking measures at each time point.

**Implicit evaluations.** I computed the proportion of “Pleasant” responses to the pictographs on trials following the different prime types, at both Time 1 and Time 2, and used these proportions as measures of implicit evaluative impressions of the primes. I then analyzed these proportions in a 2 (Time: time 1, time 2) x 2 (Prime: Francis West, control faces) x 5 (Story Condition: control, reinterpretation, negation only, affirm only, negate + affirm) mixed ANOVA, with the last factor manipulated between-participants.

All main effects and interactions in the model were significant, with the exception of the main of effect of story condition. Most critical, however, was the significant three-way interaction between time, prime, and story condition,  $F(4, 339) = 10.694, p < .001, \eta_p^2 = .112$ . Breaking down this interaction, I found that the interaction between time and prime was not significant in the control condition,  $F(1, 72) = .958, p = .331, \eta_p^2 = .013$ , but was significant in the other four (see below). Among those four conditions, there remained a significant interaction between time, prime, and story condition,  $F(3, 267) = 4.970, p = .002, \eta_p^2 = .053$ , which indicates that the highest-order interaction cannot be attributed solely to the difference of the control condition from the other four conditions.

I next constructed three orthogonal contrasts to compare the size of the time\*prime interaction in the different between-participants story conditions, setting aside the control condition: The first tested the reinterpretation and negate + affirm conditions (pooled) against the negate only and affirm only conditions (pooled), the second tested the reinterpretation condition against the negate + affirm condition, and the third tested the negate only condition against the

affirm only condition. The first contrast was significant,  $F(1,267) = 13.006, p < .001, \eta_p^2 = .046$ , suggesting that collectively, the fire rescue (reinterpretation) condition and the combined negate + affirm condition differed from the negate only and affirm only conditions in regards to changes in implicit evaluations of Francis West. However, the second contrast was not significant,  $F(1,267) = 2.575, p = .134, \eta_p^2 = .008$ , and neither was the third,  $F(1,267) = .071, p = .791, \eta_p^2 < .001$ , indicating no evidence for differences between the reinterpretation and negate + affirm conditions on the one hand, or the negate only and affirm only conditions on the other hand.

Figure 8 shows the average implicit evaluations for both prime types (Francis West and control) across time in all five story conditions.

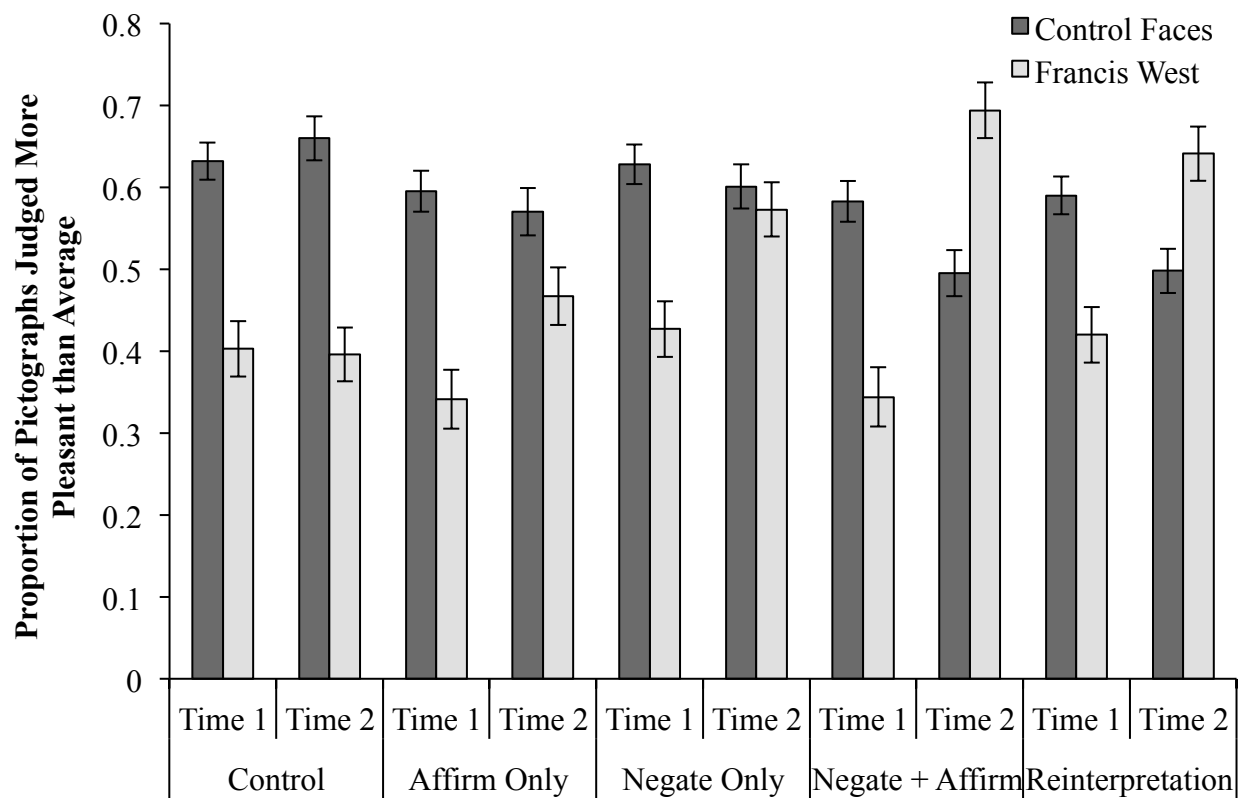


Figure 8. Mean proportion of pictographs judged as more pleasant than average in Study 7 by measurement time, story condition, and face prime. Error bars are standard errors.

The interaction between prime and time was significant in both the reinterpretation ( $F[1,71] = 24.579, p < .001, \eta_p^2 = .257$ ) and negate + affirm conditions ( $F[1,64] = 46.260, p < .001, \eta_p^2 = .420$ ). In the reinterpretation condition at Time 1, Francis West was less implicitly positive ( $M = .420, SD = .263$ ) than control faces ( $M = .590, SD = .202$ ),  $t(71) = 3.652, p < .001$ , Hedges'  $g_{av} = .717$  (using the effect size measure recommended by Lakens, 2013; this can be interpreted as similar to a between-participants Cohen's  $d$ , with correction for bias due to sample size). In the negate + affirm condition at Time 1, Francis was also less implicitly positive ( $M = .344, SD = .284$ ) than control faces ( $M = .583, SD = .180$ ),  $t(64) = 5.501, p < .001$ , Hedges'  $g_{av} = .991$ . At Time 2, implicit evaluations had reversed in both conditions. In the reinterpretation condition, Francis West was more implicitly positive ( $M = .641, SD = .251$ ) than control faces ( $M = .498, SD = .261$ ),  $t(71) = 3.056, p = .003$ , Hedges'  $g_{av} = .553$ . In the negate + affirm condition, Francis was also more implicitly positive ( $M = .694, SD = .261$ ) than control faces ( $M = .495, SD = .197$ ),  $t(64) = 4.991, p < .001$ , Hedges'  $g_{av} = .850$ . The reinterpretation and negate + affirm conditions, then, showed very similar patterns of results.

The interaction between time and prime was also significant in the negate only ( $F[1,70] = 11.568, p = .001, \eta_p^2 = .142$ ) and affirm only conditions ( $F[1,62] = 7.085, p = .010, \eta_p^2 = .103$ ). In the negate only condition at Time 1, Francis West was less implicitly positive ( $M = .427, SD = .305$ ) than control faces ( $M = .628, SD = .176$ ),  $t(70) = 4.781, p < .001$ , Hedges'  $g_{av} = .802$ . In the affirm only condition at Time 1, Francis was also less implicitly positive ( $M = .341, SD = .289$ ) than control faces ( $M = .595, SD = .221$ ),  $t(62) = 5.864, p < .001$ , Hedges'  $g_{av} = .972$ . Unlike in the reinterpretation and negate + affirm conditions, however, the implicit evaluations did not reverse at Time 2. In the negate only condition, implicit evaluations of Francis West ( $M = .573, SD = .293$ ) and the control faces ( $M = .601, SD = .213$ ) did not significantly differ at Time 2,



$t(70) = .675, p = .502$ , Hedges'  $g_{av} = .109$ . In the affirm only condition, Francis West remained marginally less implicitly positive ( $M = .467, SD = .299$ ) than control faces ( $M = .570, SD = .220$ ),  $t(62) = 1.852, p = .069$ , Hedges'  $g_{av} = .385$ .

**Explicit evaluations.** The six items measuring explicit liking of Francis West showed high reliability both at Time 1 (Cronbach's  $\alpha = .94$ ) and Time 2 (Cronbach's  $\alpha = .99$ ), and were thus merged into a single aggregate score for each point in time. I analyzed these scores in a 2 (Time: Time 1, Time 2) x 5 (Story Condition: control, reinterpretation, negation only, affirm only, negate + affirm) mixed ANOVA, with Time manipulated within-participants and Story Condition manipulated between-participants.

There was a main effect of time,  $F(1,339) = 1824.50, p < .001, \eta_p^2 = .843$ , and a main effect of condition,  $F(4,339) = 202.24, p < .001, \eta_p^2 = .943$ , but both were qualified by a significant interaction between time and condition,  $F(4,339) = 229.89, p < .001, \eta_p^2 = .731$ . At Time 1, there was no effect of story condition on liking,  $F(4,339) = .616, p = .651, \eta_p^2 = .007$ , with average liking of Francis falling significantly below the midpoint of the scale ( $M = 1.24, SD = .59, t[343] = -86.81, p < .001$ ). At Time 2, however, story condition had a significant effect on liking,  $F(4,339) = 275.86, p < .001, \eta_p^2 = .765$ .

Looking at the specific conditions, I first determined that in the control condition, there was no effect of time on explicit liking,  $t(72) = 1.445, p = .153$ , Hedges'  $g_{av} = .135$ . Next, I found that the interaction between time and condition was significant across the other four story conditions,  $F(1,267) = 110.52, p < .001, \eta_p^2 = .554$ . To investigate this interaction, similar to what I did with implicit evaluations, I constructed three orthogonal contrasts to compare the size of the time effect in the different between-participants story conditions, setting aside the control condition: The first tested the reinterpretation and negate + affirm conditions (pooled) against the

negate only and affirm only conditions (pooled), the second tested the reinterpretation condition against the negate + affirm condition, and the third tested the negate only condition against the affirm only condition.

The first contrast was significant, showing that the effect of time in the reinterpretation and negate + affirm conditions was collectively different from the effect of time in the negate only and affirm only conditions,  $F(1,267) = 307.32, p < .001, \eta_p^2 = .535$ . The second contrast was also significant,  $F(1,267) = 5.29, p = .022, \eta_p^2 = .019$ , suggesting that at the explicit level (unlike at the implicit level), the reinterpretation and negate + affirm conditions were significantly different. Finally, the third contrast was also significant,  $F(1,267) = 28.83, p < .001, \eta_p^2 = .097$ , indicating that the negate only and affirm only conditions also differed in their impact on explicit liking over time, departing from the lack of difference between these conditions on the implicit measure.

At Time 2 in the reinterpretation condition, Francis West was rated as significantly more positive ( $M = 5.89, SD = 1.59$ ) than in the control condition ( $M = 1.16, SD = .31$ ), affirm only condition ( $M = 2.62, SD = 1.46$ ), and negate only condition ( $M = 4.01, SD = .91$ ), all  $ps < .001$ . However, he was rated as less positive than in the negate + affirm condition ( $M = 6.44, SD = .78$ ), unequal variances  $t(105.99) = 2.583, p = .011$ , Hedges'  $g_s = .439$ . Next, though at Time 2 in the affirm only condition Francis West was more explicitly positive than in the control condition, unequal variances  $t(66.96) = 7.814, p < .001$ , Hedges'  $g_s = 1.336$ , he was less positive than in the negate only condition, unequal variances  $t(101.03) = 6.522, p < .001$ , Hedges'  $g_s = 1.122$ . Importantly, even though the negate only condition produced more positive explicit evaluations of Francis West at Time 2 than the affirm only condition, they were still significantly less

positive than in the reinterpretation condition, as reported above. Figure 9 shows the average explicit liking ratings of Francis West across time in each of the five story conditions.

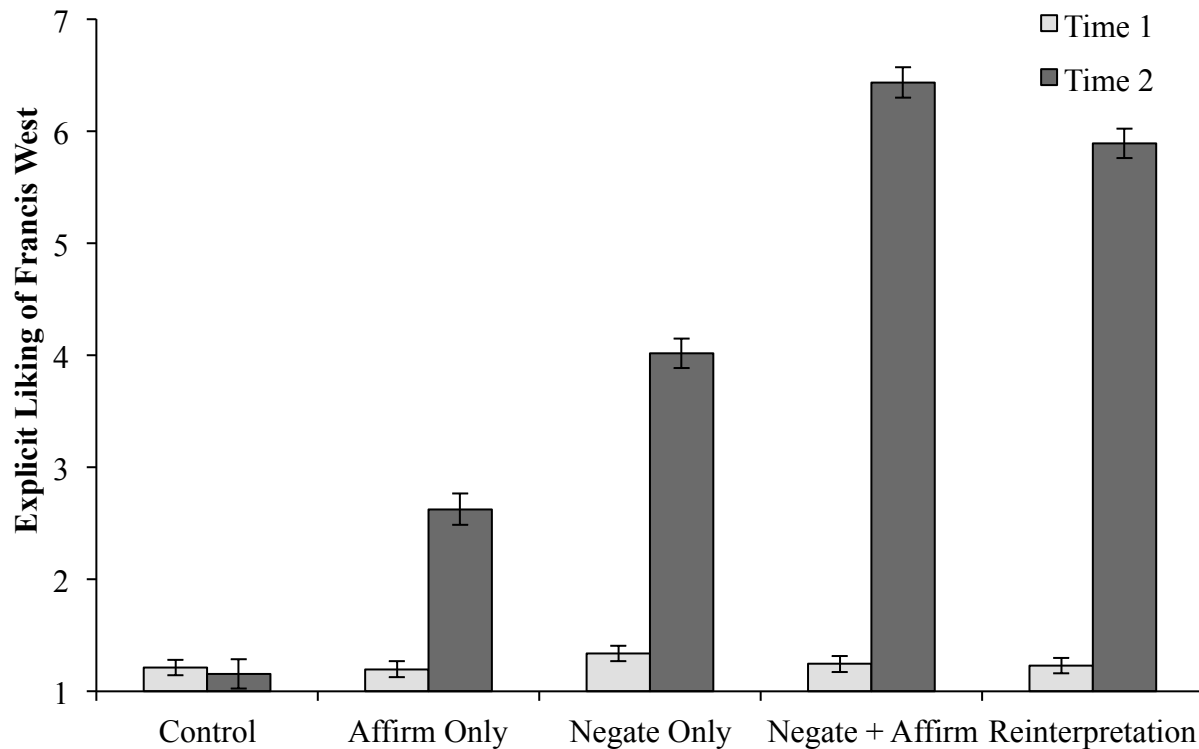


Figure 9. Mean explicit liking of Francis West in Study 7, by story condition and measurement time. Error bars are standard errors.

## Discussion

The results of Study 7 supported the proposal that the combined negation of an earlier impression and affirmation with a new, countervailing one drives effective revision in implicit impressions, to a degree that is indistinguishable from the effects of reinterpretation. By comparing the effectiveness of the reinterpretation condition to others in which negation, affirmation, and their combination are experimentally controlled, I found that only the combined negate + affirm condition was also effective in producing reversal of implicit evaluations. Importantly, this combination suggests that at a broad level, new information that both *negates*

the initial impression and *replaces* it with an updated one can update implicit responses to the person, regardless of whether those two steps are intrinsically related (as is the case with reinterpretation) or realized in a piecemeal fashion. This possibility was further supported by the lack of a significant difference in the size of the revision in implicit evaluations between the two conditions. Though the negate + affirm condition was actually more effective than the reinterpretation condition in changing *explicit* evaluations, this difference was not hypothesized and failed to replicate in the following Study 8; I thus hesitate to draw strong conclusions from that effect.

Though these data clearly show that combining negation of an earlier impression with the addition of a new one is more effective than either step in isolation, the similar degree of revision in the negate + affirm condition and the reinterpretation condition is only suggestive of an underlying commonality between them. Indeed, it is possible that they operate through different mechanisms and only happen to show similar levels of revision. In Study 8, I make use of formal processing modeling to further support the proposal that reinterpretation and negate + replace produce similar revision in implicit impressions. In doing so, I turn to the question of *how* both types of new information impact responding on the implicit measure. Does the new information actually alter the implicit evaluative impressions of Francis West, or instead does it primarily impact other processes that contribute to responding on the AMP, such as the frequency of misattribution (i.e., the likelihood that the evaluation of the prime will determine the response to the pictograph)? Although I found in Study 7 that reinterpretation and negate + replace brought about similar levels of revision overall, it is possible that reinterpretation alters the evaluative impression spontaneously activated by the target person, whereas other strategies for negating and replacing a negative impression merely alter the degree to which evaluative impressions of

the person drive responses on the implicit measure (i.e., the likelihood of misattribution on the AMP). Study 8 tested these possibilities.

### **Study 8: A Process Model of Reinterpretation**

In the previous study, I found that the level of revision in the reinterpretation and negate + affirm conditions was very similar, suggesting that reinterpretation might be just one form of a broader category of strategies combining the negation and replacement of an earlier impression to effectively reverse implicit impressions. The goals of Study 8 were twofold. First, I aimed to further test the similarity of reinterpretation and other combinations of negation and affirmation by examining the similarity of implicit evaluations in the reinterpretation and negate + affirm conditions, but in a different manner from Study 7. Whereas in Study 7 I focused on the overall proportion of positive responses on the implicit measure, in Study 8, I made use of multinomial process tree modeling to test whether the two conditions would affect the processing parameters that are fitted to those response data in a similar way. While similarity in overall levels of implicit revision was established in Study 7 using the response data on the AMP, with mathematical process modeling I can examine whether the parameters meant to represent the processing components underlying responding on the task are also similar between the two conditions after the equations are fit to the data. To the degree that the parameters are similar, the proposal that both forms of information update implicit impressions in a similar manner is corroborated. Formal process modeling thus allows for similarity between the reinterpretation and negate + affirm conditions to be assessed at another level.

This leads naturally into the second goal of Study 8: To begin testing more finely the nature of the effect that the new information in both conditions has on implicit impressions. Though it is clear that reinterpretation and negate + replace information both shift implicit

impressions, this shift tells us little about the processes of implicit evaluation that are impacted. While I hypothesize that changes to the evaluative response that is activated by the person prime is chiefly (or solely) responsible for the effects of the information, it is also possible that the likelihood of *misattributing* that response to the pictograph is also affected. Multinomial process tree modeling provides one approach to testing these possibilities.

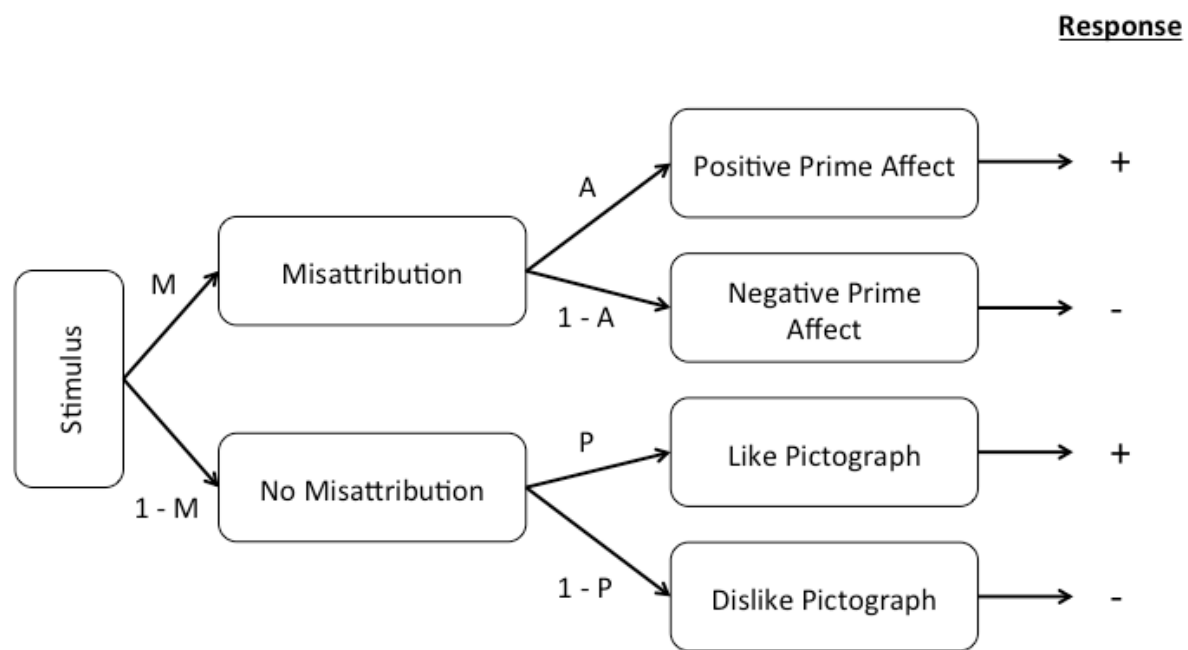
In such models, equation parameters are designed to reflect posited mental operations, and their values signify the likelihood of that operation's occurrence. The response data are then used to fit the parameters algebraically (e.g., Jacoby, 1991, Payne, 2001) or through maximum likelihood estimation (e.g., Conrey et al., 2005; Payne et al., 2010; Sherman et al., 2008). The fitted equations can then be used to generate predicted response data, with the performance of the model assessed by computing the error between the predicted and actual responses. This procedure then allows for an examination of the effect of manipulations on model parameters, which provides evidence for an effect on the underlying process that the parameter represents. The validity of the parameter's interpretation, of course, must also be assessed, usually by testing whether the parameters are affected in expected ways by manipulations during pretesting.

I made use of a previously validated multinomial process model of responding on the AMP (Payne et al., 2010). Figure 10 shows the process tree of this model. The AMP model construes responses on the AMP as driven by three processes, each captured by a unique parameter: The *A* parameter models the probability that a "pleasant" impression will be activated by the prime, the *M* parameter estimates the probability that the impression of the prime will be misattributed to the pictograph, and the *P* parameter estimates the probability that, absent misattribution, a pleasant impression will be activated by the pictograph itself. The probability of giving a "pleasant" or "unpleasant" response to each pictograph through a particular pathway in

the model is equal to the product of all of the parameters in that pathway, and the total probability of giving each type of response is equal to the sum of all branches that result in that type of response. The total probabilities of giving a “pleasant” or “unpleasant” response are thus represented by the equations:

$$p(\text{Pleasant response}) = M \cdot A + (1-M) \cdot P \quad (1)$$

$$p(\text{Unpleasant response}) = M \cdot (1-A) + (1-M) \cdot (1-P) \quad (2)$$



*Figure 10.* Multinomial process tree for the AMP model in Study 8 for pleasant (+) and unpleasant (-) responses, adapted from Payne et al. (2010).

This decomposition of AMP responses into three underlying processes has previously fit AMP data well (Payne et al., 2010), making it suitable for lending support for or against the proposal that the reinterpretation and negate + affirm conditions would show similarities in their underlying processes. For example, even if revision in overall AMP responses is similar between the two conditions, it might be the case that reinterpretation produces a smaller change in the activation of a pleasant response to the Francis West prime relative to the negate + affirm condition (*A* parameter), but an increase in the likelihood of misattribution (*M* parameter), resulting in similar levels of revision in the overall responses. My prediction, however, was that just as with the overall pattern of responses on the AMP, the parameters capturing subsidiary processes would also not significantly differ between the reinterpretation and negate + affirm conditions. In contrast, I expected that both of these conditions would differ in some way from the control story condition (in which Francis performs one more negative action, consistent with the earlier story). While I thought it likely that the *A* parameter would show a substantial difference between the control condition and the others at Time 2 (being lower in the control condition, indicating a lower probability of activating a positive impression of Francis in that condition), I did not have strong expectations about whether the *M* parameter (likelihood of misattribution) would vary as well.

## **Method**

**Participants and design.** Three hundred seventy-seven undergraduates participated in the study in return for partial course credit. Of these, only partial data was recorded on one of the implicit measures for 26 participants due to a server error, and I determined prior to looking at the data to exclude these participants from all analyses. This left full data for 351 participants. Of



these, 7 were excluded for only pressing a single response key on at least one of the implicit measures and a further 34 for familiarity with Mandarin or Cantonese (Payne et al., 2005) in accordance with protocol. This left a final sample size of 310 participants (105 men, 205 women,  $M_{\text{Age}} = 18.60$  years,  $SD = 1.61$ ). I initially aimed *a priori* to collect enough data to achieve 75 participants per each of three between-participants conditions after planned exclusions (225 total), but increased this to 100 per cell (prior to all analyses) due to an increased supply of participants. Each participant was randomly assigned to either the control, reinterpretation, or negate + affirm condition.

**Materials and procedure.** The materials and procedure were identical to those used in Study 7, with the following exceptions. First, the *negate only* and *affirm only* conditions were not included. Second, each AMP was modified to accommodate the multinomial modeling analysis. Fitting the AMP process model to AMP data requires that half of the trials present pictographs that are known to be relatively less pleasant than average, and that the other half of the trials present pictographs that are known to be relatively more pleasant than average. To generate two sets of pictographs that differ in known pleasantness, I recruited 40 participants from Mechanical Turk to complete a single AMP, consisting of 200 trials of only pictograph targets (no primes were included). For each of the 200 pictographs, I calculated the proportion of participants in the sample who responded that the pictograph was more pleasant than average, and then I sorted the pictographs according to their average pleasantness. The 60 pictographs with the highest average pleasantness became the “pleasant” set ( $M = .67$ ,  $SD = .04$ , range = .63 - .78), and the 60 pictographs with the lowest average pleasantness become the “unpleasant” set ( $M = .48$ ,  $SD = .04$ , range = .38 - .53), in that they were *relatively* less pleasant than the others. Because I planned to administer the AMP twice in the main experiment (at Time 1 and Time 2), I randomly

assigned 30 pictographs from the “pleasant” group and 30 pictographs from the “unpleasant” group to one set and the remaining pictographs from each group to the other set, creating two final sets of 60 pictographs (each half pleasant and half unpleasant). For each participant, one set of 60 pictographs was used during the Time 1 AMP, and the other set was used during the Time 2 AMP, order counterbalanced. Each AMP in this experiment thus consisted of 60 trials, with 30 trials including Francis West as the prime, and 30 including a control face as the prime. This resulted in 15 trials of each of the following types being presented during each AMP: Francis West + pleasant pictograph, Francis West + unpleasant pictograph, control face + pleasant pictograph, control face + unpleasant pictograph.

With the exception of these differences, the rest of the materials and procedure of this experiment proceeded in an identical fashion to Study 7.

## **Results**

**Data preprocessing.** For both the Time 1 and Time 2 AMPs, for each participant I calculated the proportion of trials on which the pictograph was judged to be more pleasant than average following the Francis West prime image, and a similar proportion for trials following control face images, separately for the “pleasant” and “unpleasant” pictographs, yielding 8 AMP measures for each participant within a 2 (Time: Time 1, Time 2) x 2 (Prime: Francis West, Control) x 2 (Pictograph Valence: Pleasant, Unpleasant) design. Story Condition (Control, Reinterpretation, Negate + Affirm) was the sole between-participants factor in the design. Explicit evaluations of Francis West at Time 1 and Time 2 were computed by averaging the six liking measures at each time point, as done in Study 1.

**Implicit evaluations.** I conducted a factorial mixed ANOVA on implicit pleasantness, fully crossing the four factors mentioned above (time, prime, pictograph valence, and story

condition). A significant interaction between time, prime, and story condition obtained,  $F(2,307) = 14.44, p < .001, \eta_p^2 = .086$ . This three-way effect was not moderated by pictograph valence,  $F(2,307) = 1.77, p = .173, \eta_p^2 = .011$ . There was a main effect of pictograph valence,  $F(1,307) = 48.59, p < .01, \eta_p^2 = .137$ , such that pleasant pictographs were indeed judged as more pleasant ( $M = .53, SD = .14$ ) than unpleasant pictographs ( $M = .47, SD = .14$ ). This was moderated by time,  $F(1,307) = 6.51, p = .011, \eta_p^2 = .021$ , such that this average pleasantness difference between the pictograph types was larger at Time 2 ( $M_{\text{Diff}} = .072, SE = .01$ ) than at Time 1 ( $M_{\text{Diff}} = .047, SE = .009$ ), but the difference was significant at both Time 1,  $F(1,307) = 25.32, p < .001, \eta_p^2 = .076$ , and Time 2,  $F(1,307) = 48.27, p < .001, \eta_p^2 = .136$ . Pictograph valence did not significantly moderate any other effects, all  $F_s < 2.38$ , all  $p_s > .093$ .

Turning back to the interaction of between time, prime, and story condition, I next tested two orthogonal contrasts, the first comparing the size of the interaction between time and prime in the control condition to the size of that interaction in the reinterpretation and negate + affirm conditions (pooled), and the second comparing the size of the interaction between time and prime in the reinterpretation condition to the size of that interaction in the negate + affirm condition. The first contrast was significant,  $F(1,307) = 27.29, p < .001, \eta_p^2 = .082$ , suggesting that the control condition did indeed differ from the other two (reinterpretation and negate + affirm). The second contrast, however, was not significant,  $F(1,307) = 1.40, p = .238, \eta_p^2 = .005$ , suggesting that the reinterpretation and negate + affirm conditions were similar in the size of the interaction between time and prime, replicating the null effect observed in Study 7. Figure 11 displays the mean proportion of pictographs judged to be more pleasant than average in each cell of the interaction between time, prime-type, and story condition.

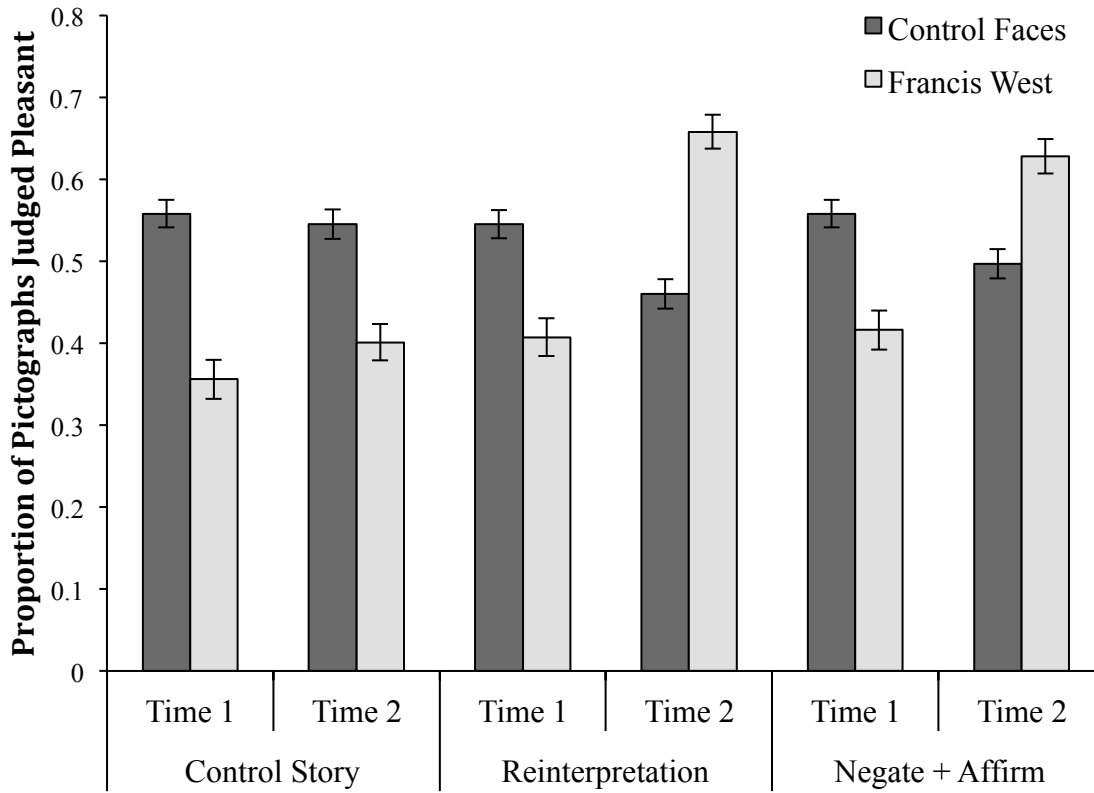


Figure 11. Mean proportion of pictographs judged as more pleasant than average in Study 8 by measurement time, story condition, and face prime. Error bars are standard errors.

In the control story condition, there was an interaction between time and prime,  $F(1,99) = 5.45, p = .022, \eta_p^2 = .052$ . Francis was significantly less pleasant than control faces at Time 1 (Francis:  $M = .36, SD = .25$ , Control:  $M = .56, SD = .17$ ),  $t(99) = 6.75, p < .001$ , Hedges'  $g_{av} = .940$ . His negativity relative to the control faces was somewhat lower at Time 2, but remained significant (Francis:  $M = .40, SD = .24$ , Control:  $M = .55, SD = .20$ ),  $t(99) = 4.67, p < .001$ , Hedges'  $g_{av} = .647$ .

In the reinterpretation condition, time significantly interacted with prime,  $F(1,107) = 57.55, p < .001, \eta_p^2 = .350$ . Francis was significantly less pleasant than control faces at Time 1 (Francis:  $M = .41, SD = .24$ , Control:  $M = .55, SD = .19$ ),  $t(107) = 4.41, p < .001$ , Hedges'  $g_{av} =$

.640, but he was significantly more pleasant than control faces at Time 2 (Francis:  $M = .66$ ,  $SD = .20$ , Control:  $M = .46$ ,  $SD = .17$ ),  $t(1,307) = 7.20$ ,  $p < .001$ , Hedges'  $g_{av} = 1.049$ . Thus, the reinterpretation information produced a significant reversal in implicit evaluations from Time 1 to Time 2, replicating the result obtained in Study 7.

In the negate + affirm condition, there was also a significant interaction between time and prime,  $F(1,101) = 42.82$ ,  $p < .001$ ,  $\eta_p^2 = .298$ . Francis was significantly less pleasant than control faces at Time 1 (Francis:  $M = .42$ ,  $SD = .22$ , Control:  $M = .56$ ,  $SD = .16$ ),  $t(101) = 5.44$ ,  $p < .001$ , Hedges'  $g_{av} = .731$ , but he was significantly more pleasant than control faces at Time 2 (Francis:  $M = .63$ ,  $SD = .20$ , Control:  $M = .50$ ,  $SD = .18$ ),  $t(101) = 4.80$ ,  $p < .001$ , Hedges'  $g_{av} = .687$ . Thus, there was a significant reversal of implicit evaluations from Time 1 to Time 2 in the negate + affirm condition, which was also consistent with the findings of Study 7.

**Explicit evaluations.** The six items measuring explicit liking of Francis West showed high reliability both at Time 1 (Cronbach's  $\alpha = .84$ ) and Time 2 (Cronbach's  $\alpha = .99$ ), and were thus merged into a single aggregate score for each point in time. I analyzed these scores in a 2 (Time: Time 1, Time 2)  $\times$  5 (Story Condition: control, reinterpretation, negate + affirm) mixed ANOVA, with Time manipulated within-participants and Story Condition manipulated between-participants.

There was a significant main effect of time,  $F(1,307) = 3869.54$ ,  $p < .001$ ,  $\eta_p^2 = .926$ , and story condition,  $F(2,307) = 1198.39$ ,  $p < .001$ ,  $\eta_p^2 = .886$ , which were both qualified by an interaction between time and story condition,  $F(2,307) = 1001.15$ ,  $p < .001$ ,  $\eta_p^2 = .867$ . I next compared the time effect in the different conditions using two orthogonal contrasts, the first testing the size of this effect in the control condition vs. the reinterpretation and negate + affirm conditions (pooled), and the second comparing the size of the time effect in the reinterpretation

condition vs. the negate + affirm condition. The first contrast was significant,  $F(1,307) = 2000.23, p < .001, \eta_p^2 = .867$ , suggesting that the effect of Time in the control condition did indeed differ from the effect of Time in other two. The second contrast was not significant,  $F(1,307) = .509, p = .476, \eta_p^2 = .002$ . This indicates that the explicit evaluations of Francis West changed equally from Time 1 to Time 2 in the reinterpretation and negate + affirm conditions, paralleling the results on the implicit measure (and not replicating the explicit attitude difference found between these conditions in Study 7). At Time 2, Francis West was rated as more positive in the reinterpretation condition ( $M = 6.25, SD = .74$ ) than in the control condition ( $M = 1.29, SD = .53$ ), unequal variances  $t(194.501) = 56.045, p < .001$ , Hedges'  $g_s = 7.749$ , but he was not rated any differently in the reinterpretation condition than in the negate + affirm condition ( $M = 6.25, SD = .77$ ),  $t(208) = .046, p = .963$ , Hedges'  $g_s = .006$ .

**AMP process model.** I next turned to fitting the AMP process model to the data obtained from the Time 1 and Time 2 AMPs (Payne et al., 2010). Because each response given by a participant falls into one cell of a 3 (Story Condition: control, reinterpretation, negate + affirm) x 2 (Time: Time 1, Time 2) x 2 (Prime: Francis, control) x 2 (Pictograph Valence: Pleasant, Unpleasant) design, there are 24 independent response categories for model fitting.<sup>1</sup> The AMP model specifies three parameters that can be permitted to vary across response categories: A (probability of activation of a pleasant response to the prime), M (probability of misattributing the response produced by the prime to the pictograph), and P (probability of activation of a pleasant response to the pictograph). Each parameter can also be understood as 1 – the probability of the opposite; for example, A can be understood as one minus the probability of activation of an unpleasant response to the prime.

Because the model has degrees of freedom equal to the number of independent response categories (24) minus the number of model parameters, it is not possible to allow all 3 parameters to vary within each response category, as this would produce a model with negative degrees of freedom ( $24 - 72 = -48$ ). Thus, I specified the following *a priori* model:

**Parameter A.** A is free to vary across cells of the story condition, time, and prime factors, but not pictograph valence (this adds 12 A parameters to the model).

**Parameter M.** M is free to vary across cells of the story condition and time factors, but not prime or pictograph valence (This adds 6 M parameters to the model).

**Parameter P.** P is free to vary across cells of the pictograph valence factor, but no others (This adds 2 P parameters to the model).

The model thus assumes that the likelihood that a pictograph *itself* will activate a pleasant or unpleasant response depends only on the valence group that the pictograph belongs to (the unpleasant pictographs set or the pleasant pictographs set), consistent with specifications of this parameter in prior work (Payne et al., 2010). Also consistent with prior work, it was assumed that the likelihood of a prime activating a pleasant response (A), or the likelihood of misattributing the response triggered by the prime to the pictograph (M), would not vary with the pictograph valence. I allowed M to vary by time and story condition, but assumed that the type of prime on each trial (Francis or control) would not affect the probability of misattribution, also similar to choices made in prior work (Payne et al., 2010). Finally, A was given the most leeway to vary, as I assumed that the probability that a prime would produce a pleasant response would depend on the type of prime (Francis vs. control), time, and story condition. This model thus had a total of 20 free parameters, resulting in  $24 - 20 = 4$  degrees of freedom. The responses of all participants in the sample within each response category were pooled to create a single set of

counts, and the model was fit with the “mpt” package in R (Wickelmaier, 2011), using the EM algorithm (Hu & Batchelder, 1994).

**Model results.** Overall model fit was satisfactory (as indicated by a *non-significant*  $p$ -value),  $G^2 = 8.689$ ,  $p = 0.069$ . Table 3 displays the values fit to each of the 20 model parameters.<sup>2</sup>

Table 3. *AMP multinomial process model parameter estimates, Study 8*

Parameter	Control Condition		Reinterpretation Condition		Negate + Affirm Condition	
	Francis Prime	Control Prime	Francis Prime	Control Prime	Francis Prime	Control Prime
Time 1						
A	.244	.591	.350	.563	.365	.582
M		.582		.649		.656
Time 2						
A	.246	.591	.892	.374	.752	.481
M		.419		.383		.482
	More Pleasant Pictographs			Less Pleasant Pictographs		
P		.580			.450	

*Note.* Parameter A = the probability of activation of a “pleasant” response to the prime, and 1 – the probability of activation of an “unpleasant” response to the prime. Parameter M = the probability of misattributing the response produced by the prime to the pictograph, and 1 – the probability of not misattributing the response produced by the prime to the pictograph, in which case the model assumes that responses are driven by parameter P. Parameter P = the probability of activation of a “pleasant” response to the pictograph, and 1 – the probability of activation of an “unpleasant” response to the pictograph.



**Model tests.** The differences between parameters can be tested for significance by refitting the model while constraining the parameters to be equal, and then taking the difference of the  $G^2$  value of the old model and the new model. This difference is tested against the null value of zero in a one-tailed chi-square test with degrees of freedom equal to the difference in number of parameters between the two models (the test is one-tailed because the  $G^2$  value cannot be smaller in the more constrained model).

*Tests of P.* I first tested whether the probability of activating a “pleasant” response to a pictograph from the more pleasant set (.58) was significantly greater than the probability of activating a “pleasant” response to a pictograph from the less pleasant set (.45). As expected, this was the case; a model constraining the two P parameters to be equal had significantly worse fit than the original model,  $\chi^2(1) = 142.21, p < .001$ . Though this finding amounts to a manipulation check, it was important for suggesting that the model was performing reasonably.

*Tests of M.* Next, I tested whether the probability of misattribution varies across time and story conditions; examining the parameter estimates in Table 1, there is some suggestion of a decrease in the likelihood of misattribution from Time 1 to Time 2 in all story conditions. I thus fit a model constraining M to be equal in all six cells, and found that the resulting constrained model was not significantly worse than the original model,  $\chi^2(5) = 7.10, p = .214$ .

*Tests of A.* Finding no evidence for strong differences in misattribution across time and story conditions, I next turned to the probability of activation of a “pleasant” response to the primes (parameter A). First, I tested whether the probability of activation of a pleasant response to Francis West was lower than the probability of activation of a pleasant response to the control primes at Time 1 in each story condition. This was indeed the case: The A parameter was

significantly lower for Francis West at Time 1 relative to the control faces in the control condition (Francis: .244, control faces: .591),  $\chi^2(1) = 248.94, p < .001$ , the reinterpretation condition (Francis: .350, control faces: .563),  $\chi^2(1) = 124.15, p < .001$ , and the negate + affirm condition (Francis: .365, control faces: .582),  $\chi^2(1) = 124.98, p < .001$ .

At Time 2, I expected that the A parameter would remain lower for Francis relative to the control face primes in the control story condition, but would become higher for Francis relative to the control face primes in the reinterpretation and negate + affirm conditions. First, I tested whether the model would be significantly worse if all of the A parameters for Francis West were constrained to be equal at Time 2. This did significantly worsen model fit, suggesting that the likelihood of activating a pleasant response to Francis West at Time 2 varied across conditions,  $\chi^2(2) = 28.55, p < .001$ . This was not the case for the control faces, as constraining the A parameters for the control faces to be equal at Time 2 across story conditions did not significantly worsen model fit,  $\chi^2(2) = 3.08, p = .214$ .

In the control story condition, the A parameter did indeed remain lower for Francis (.246) relative to control face primes (.591) at Time 2,  $\chi^2(1) = 126.67, p < .001$ . Also in line with my predictions, in the reinterpretation condition at Time 2, a pleasant response was more likely to be activated by the Francis West prime than by control face primes: the A parameter for Francis West (.892) was significantly higher than the A parameter for control faces (.374),  $\chi^2(1) = 262.28, p < .001$ . Finally, as expected, a reversal in the relative size of the A parameter also occurred in the negate + affirm condition at Time 2, with the A parameter being significantly larger for Francis West (.752) than for control primes (.481),  $\chi^2(1) = 106.93, p < .001$ .

## **Discussion**

The results of Study 8 were thoroughly consistent with the proposal that information that prompts negation of an earlier impression and its replacement with a new impression, whether through reinterpretation or discrete steps, reverses implicit evaluations in a similar pattern. The similar degree of revision between the reinterpretation and negate + affirm conditions in Study 7 was replicated here, and the application of an established multinomial processing model to the data (Payne et al., 2010) showed support for the equivalence in underlying processes on the implicit measure between the two conditions. Both the reinterpretation and negate + affirm conditions, however, differed in an anticipated way from the control condition, in which Francis remained negative from Time 1 to Time 2: The  $A$  parameter continued to suggest the low likelihood of activation of a positive impression of Francis West (and the high likelihood of activation of a negative impression) at Time 2 in the control condition. Furthermore, the value of the  $A$  parameter significantly differed between the control condition and the other two conditions, whereas the value of the  $M$  parameter did not.

This finding – that the impact of the reinterpretation and negate + affirm conditions on implicit impressions of Francis West is driven in both cases by changes in the impression activated by the prime, rather than the likelihood of misattribution – supports the conclusion that the new information alters the evaluative impression of the target that is spontaneously primed, rather than other processes that impact responses. In the AMP model, another such process is the likelihood of misattribution, modeled by the  $M$  parameter, which did not vary across time or information conditions. It was thus not the case that reinterpretation (or negate + affirm) increased the likelihood of misattribution. If anything, the likelihood of misattribution seemed to *decrease* from Time 1 to Time 2 across conditions, though not significantly so. The results of Study 8, then, suggest that the effectiveness of the forms of new information examined here for

updating implicit impressions results in an altered evaluative impression of the target, regardless of whether the negation and replacement are achieved through reinterpretation.

Having now developed and tested a model of the constituent processes that may occur not only during reinterpretation, but also in other potentially successful routes of revision within a broader class of “negate + affirm” strategies, this chapter next turns to the question of how effectively reinterpretation can overturn an implicit first impression after a greater amount of time has passed since initial formation – a topic of critical importance for establishing the generalizability of reinterpretation as a mechanism of implicit updating.

## Study 9: Revision After Delay<sup>13</sup>

### Abstract

People are adept at forming impressions of others, but how easily can impressions be updated? Although implicit first impressions have been characterized as difficult to overturn, recent work shows that they can be reversed through reinterpretation of earlier learning. However, such reversal has been demonstrated only in the same experimental session in which the impression formed, suggesting that implicit updating might be possible only within a brief temporal window, before impressions are consolidated and when memory about the initial information is strongest. Implicit impressions may be unable to be revised when reinterpreting details are learned later, due to memory consolidation or forgetting of the details to be reinterpreted. This study tested whether implicit first impressions can be reversed through reinterpretation after a two-day delay following the initial formation. Results showed that implicit revision emerged after the delay, even among those with poor explicit recall or who were not cued to recall. We discuss implications for theory on impression formation and updating.

Keywords: implicit evaluation; first impressions; AMP; reinterpretation; attitudes; recall

---

<sup>13</sup> Published as: Mann, T. C., & Ferguson, M. J. (2017). Reversing implicit first impressions through reinterpretation after a two-day delay. *Journal of Experimental Social Psychology*, 68, 122–127. <http://doi.org/10.1016/j.jesp.2016.06.004> © Elsevier, 2017.

People are adept at making inferences about other people based on even minimal information, such as brief descriptions of single behaviors (Todorov & Uleman, 2002), facial characteristics (Rule et al., 2010), and group membership (Devine, 1989). These impressions have consequences for judgments and interpersonal behavior, and persist even when the initial information can no longer be recalled (e.g., McCarthy & Skowronski, 2011; Todorov & Uleman, 2002). But what happens when first impressions are wrong? Is it true that we “never get a second chance to make a first impression”? First impressions that are activated implicitly (unintentionally), rather than explicitly (intentionally), indeed have been repeatedly portrayed as more difficult to undo (Gregg et al., 2006; McConnell & Rydell, 2014; Wilson, Lindsey, & Schooler, 2000). Explanations for this dissociation vary, but a common proposal is that implicit evaluations are primarily driven by slow-changing associations, which are less sensitive to propositional processing – suggesting that first impressions that have been explicitly rejected might continue to implicitly guide behavior (e.g., Rydell & McConnell, 2006).

Recent work, however, indicates that implicit impressions can be rapidly reversed. First, theoretical models and data have increasingly supported the idea that propositional thinking *can* impact implicit processes (e.g., De Houwer, 2014; Gawronski & Bodenhausen, 2006; Wyer, 2010, 2016), questioning the view that implicit impressions must be inherently slow to change. For example, although approach-avoidance behaviors can impact implicit evaluations of stimuli (e.g., Kawakami, Phills, Steele, & Dovidio, 2007; Woud, Maas, Becker, & Rinck, 2013), merely *instructing* participants about approach-avoidance contingencies can similarly impact implicit evaluations (Van Dessel, De Houwer, Gast, & Smith, 2015; Van Dessel, De Houwer, Gast, Smith, & Schryver, 2016). In the context of impression change, these ideas suggest that revision of implicit impressions may be strongest when new information is subjectively assessed as more

diagnostic or important (though even propositional information that is seemingly low in diagnosticity may sometimes drive change; Van Dessel et al., 2016). In support of this possibility, Cone and Ferguson (2015) asked participants to read 100 behavioral statements about a novel person that indicated his overall goodness, and formed implicit positive impressions of him. After then learning an additional behavior that was extremely negative (e.g., that he had molested children), participants showed robust revision, switching to a strongly negative implicit impression of him. Furthermore, the perceived offensiveness of the single negative action predicted the extent of revision, and this effect was mediated by the perceived diagnosticity of the actions (i.e., how much the behavior was assumed to reflect the target's true character).

It may be harder, however, to revise an initially *negative* implicit impression to a positive one in this way, in that negative information is viewed as particularly diagnostic (Skowronski & Carlston, 1989). In their work, Cone and Ferguson (2015) in fact did not find reversal after participants learned that an initially negative person also performed an extremely positive behavior (donating a kidney to a sick child). However, another set of studies (Mann & Ferguson, 2015) showed that revising even an initial *negative* implicit impression is possible. Namely, participants who learned new information that changed the evaluative meaning of the initial information showed strong implicit revision: when participants read about a man who broke into and damaged his neighbor's homes, the ensuing negative implicit impression was reversed by the discovery that he was actually rescuing children from a fire. The degree of updating was predicted by participants' self-reported extent of reinterpretation. Other work has also shown that when new information is presented that updates previously learned details, implicit evaluations are likely to be revised when participants can elaborate on the earlier information

(Wyer, 2010). This suggests that reinterpretation may be a powerful route through which even negative implicit first impressions can be reversed.

### **A Brief Window of Opportunity for Implicit Revision through Reinterpretation?**

One major limitation of existing work on reinterpretation (Mann & Ferguson, 2015; Wyer, 2010), as well as all other work showing implicit revision of first impressions toward novel targets (Cone & Ferguson, 2015; Peters & Gawronski, 2011; Wyer, 2016), is that in all cases the new evidence is presented mere minutes after the first impressions have formed. Thus, even though new information that reinterpreted the meaning of a first impression was successful in “undoing” implicit first impressions in a single lab session (Mann & Ferguson, 2015), it is possible that implicit revision can occur *only* within this kind of short temporal window (see Peters & Gawronski, 2011), which would considerably limit the circumstances in which implicit impressions can be updated. Why might revision through reinterpretation be possible only within such a brief temporal window? There are at least two lines of work that are relevant. First, as time elapses after the first impression has formed, memory for the details of the initial information should subside, potentially undermining revision. The details underlying one’s initial impression are often less impactful on judgment over time as more information about a person is learned, with such judgments about the subject of an impression becoming increasingly based on trait abstractions more than recall of specific behaviors (e.g., Klein, Loftus, Trafton, & Fuhrman, 1992; Sherman & Klein, 1994; see also Hastie & Park, 1986). There is sometimes even stronger recall for behaviors that *contradict* one’s overall impression than for behaviors that are consistent with it (e.g., Babey, Queller, & Klein, 1998; Hastie & Kumar, 1979), as the efficiency gains of reliance on schemas renders continued use of memories for consistent behaviors less necessary. Indeed, first impressions persist and continue to impact how people



respond to others even after the initial behavioral information is forgotten (Castelli et al., 2004; McCarthy & Skowronski, 2011; Todorov & Uleman, 2002).

Efficiency from abstraction, however, may have a cost. In particular, *revision* of impressions through reinterpretation may require recall of the specific behaviors that led to their formation in the first place (Wyer, 2010, 2016). Precisely because the subjective diagnosticity of new information is crucially important in predicting the impact of that information on impressions (Skowronski & Carlston, 1989), people may be skeptical of new information that contradicts an earlier impression if they cannot recall its initial source. If earlier source memories have been forgotten but a person nonetheless retains the corresponding impression they formed on the basis of those memories, how can one compare the strength of the original and new information? This should be particularly difficult when the new information hinges on a reinterpretation of earlier details. Following this logic, reinterpretation may lead to updating only when one can connect the new details to the initial information. If we find this to be true in the present work, it would reveal an important limitation of trait abstraction in impression formation: a subsequent impairment in the capacity for revision, perhaps through reinterpretation in particular.

A second line of work suggesting that revision may be unlikely after time has elapsed concerns memory consolidation. Implicit impressions should consolidate over time, such that the representations may become less susceptible to interference through processes occurring over multiple timescales (Dudai, 2004; McGaugh, 2000). This would suggest that implicit first impressions become more difficult to revise once they have consolidated. Notably, this remains untested, as all research to date examining updating of implicit first impressions of novel targets has demonstrated such change before memory consolidation would be expected (i.e., within a

30-minute experiment; e.g., Cone & Ferguson, 2015; Mann & Ferguson, 2015; Peters & Gawronski, 2011; Wyer, 2010, 2016). Finding that even information that directly bears on – and reverses – the evaluative implications of earlier learning fails to update implicit evaluations after two days would expand on recent findings that a delay of even minutes makes effective updating of implicit impressions less likely (Peters & Gawronski, 2011; see also Zanon, De Houwer, Gast, & Smith, 2009).

### **Implicit Revision After A Delay**

Although some work suggests that implicit revision may be unlikely after a delay of multiple days, other work suggests that it might emerge. Namely, implicit evaluative impressions may be more reflective of recent rather than older experiences (Castelli, Carraro, Gawronski, & Gava, 2010; cf. Dunham, Baron, & Banaji, 2008; Rudman, Phelan, & Heppen, 2007; Zanon et al., 2009). For example, Castelli and colleagues (2010) found that participants' self-reported recent (versus childhood) experiences, behaviors, and feelings about religion predicted their implicit evaluations of religion.

Other research in cognitive psychology and cognitive neuroscience research shows that consolidation is not entirely unidirectional. A limitation of extinction and counterconditioning training for aversive memories is that once such initial memories are consolidated, new countervailing learning often is encoded into a separate memory trace, and thus does not replace the earlier memory (e.g., Bouton, 1994; see also Gawronski & Cesario, 2013). Recent work on reconsolidation, however, has shown that reactivation of memories produces instability that enables modification (e.g., Agren et al., 2012; Lane et al., 2015; Schiller, Kanen, LeDoux, Monfils, & Phelps, 2013; Schiller et al., 2010). Schiller and colleagues (2010), for example, found that only when participants were reminded of a conditioned stimulus 10 minutes prior to

extinction training did the extinguished response fail to reappear a day or even a year later. The plasticity of consolidated memories after reminders thus opens the door to updating even after considerable time has passed. Furthermore, it may be the case that reinterpretation is particularly likely to result in the new information being integrated into the mental representation of the initial impression, rather than result in a separate, contextualized representation (Gawronski & Cesario, 2013), given the direct relevance of the new information to the old.<sup>14</sup>

### **Overview of Current Work**

In the current study, we tested whether implicit evaluative impressions of a novel person could be reversed by new information learned two days after formation. We led participants to form a negative impression of a target, and then introduced (after two days) new details that either did or did not reframe those actions as positive. We measured implicit evaluations of the target after both sets of information.

We also examined the role of recall in revision through reinterpretation after delay. In order for revision to ensue, perhaps especially after consolidation, it may be necessary to be able to explicitly recall (and reconsider) a considerable amount of the initial information (Wyer, 2010, 2016), or to at least be prompted to remember the details. In contrast, if implicit revision is robust, it may emerge regardless of explicit memory for the details of the initial information. To test this question, we manipulated whether participants completed a memory quiz before they received the new information in the second session.

---

<sup>14</sup> We thank a reviewer for bringing this argument to our attention.

## Method

### Participants

Four hundred seventy-two participants were recruited from Mechanical Turk (mturk.com) in exchange for \$1.00 for completing the first session and \$0.50 for completing the second session. We determined through a power analysis to recruit enough participants to achieve a sample of approximately 260 participants at Time 2 for 80%+ power to detect a medium-small (Cohen's  $d = .25$ ) four-way mixed interaction effect.<sup>15</sup> Participants were recruited in batches of 50-117 over approximately one month until we exceeded our target Time 2 sample size (this was our stopping rule). These small batches allowed us to monitor attrition and ensure that all participants in a batch would finish the first session within hours of each other, making it possible for the mass invitation for session 2 to be sent approximately two days later for all participants. Five participants failed to complete the first session, and partial data from 2 were lost due to server error, so they were excluded from all analyses. Of those who completed Session 1, 62% (289 participants) completed the second session. Data from 5 participants fluent in Mandarin or Cantonese, and 14 participants who used solely one response key on every trial of at least one administration of the implicit measure (Payne et al., 2005), were excluded from all analyses *a priori*. This left a final sample of 270 participants (43% women,  $M_{\text{age}} = 32$  years; 81.9% white, 2.2% Hispanic or Latino, 7% Black, 6.7% Asian, 1.9% other race, .4% race not given).

### Session One

---

<sup>15</sup> We believe that in hindsight, we performed this power analysis incorrectly; however, the sample size is consistent with prior studies using this same paradigm (Mann & Ferguson, 2015), and Westfall's (2015) PANGAEA power analysis tool (<http://jakewestfall.org/pangea/>) suggests that the power for this 4-way design with a total sample of 260 participants is estimated at 87.2% under default variance component assumptions, which is above our target of 80% power.

During the first session, participants read a story, presented across 26 screens, in which a man named Francis West broke into the homes of two of his neighbors and caused extensive destruction. Participants proceeded through the story at their own pace, and read a variety of details about Francis's actions, such as that he broke down a door, threw a pot of water all over a young girl's computer, and removed precious things from the bedrooms. This story was designed to induce a strong negative impression of Francis (see the Supplemental Material). A photograph of a man labeled "Francis West" was presented on each screen. For each participant, one photograph of a man from a bank of 11 used in prior research (Mann & Ferguson, 2015; Minear & Park, 2004) was randomly selected for this purpose, with the other 10 serving as control stimuli during the implicit measures.

After reading the Francis West story, participants completed the first Affect Misattribution Procedure (AMP; Payne, Cheng, Govorun, & Stewart, 2005). On each of 40 trials, participants were presented with a prime image for 75 ms (Francis West on 20 trials, control faces on 20 trials). After a 125 ms blank interval, a Chinese pictograph was then presented for 100 ms before being replaced by a black and white pattern mask. The participants were instructed to judge whether each pictograph was more or less pleasant than the average pictograph, while attempting to not allow the prime stimuli to have any impact on these decisions. The extent to which the primes are systematically associated with pleasantness judgments of the pictographs across trials despite these instructions thus reveals unintentional evaluative processing of the primes (Payne et al., 2005, 2013). The 40 pictographs were drawn from one of two sets (counterbalanced across participants). (Counterbalance condition produced no significant effects, and will not be discussed further.)

Following the first AMP, participants responded to six questions assessing explicit attitudes toward Francis West. They were first asked, “How likable is Francis West?” on a scale from 1 (*very unlikable*) to 7 (*very likable*). Then, they were asked to place Francis on 1-7 scales from *bad* to *good*, *mean* to *pleasant*, *disagreeable* to *agreeable*, *uncaring* to *caring*, and *cruel* to *kind*, in random order. Finally, we collected their email address so as to contact them for the second session (all but 6 provided one), and compensated them \$1.00.

## **Session Two**

Approximately 48 hours after the end of their first session, participants were emailed a link inviting them to participate in the second session, for which they would receive a bonus (\$0.50) upon completion. They were given 24 hours to participate, with a reminder email after 12 hours (mean interval between sessions = 55.30 hours,  $SD = 7.27$ , range 47.39 to 73.76). (The size of the delay did not impact any of the results.) Participants were randomly assigned to either complete a quiz assessing their recall of the events described in the first session before learning a final piece of information about Francis, or to skip directly to learning that final information.

Those in the quiz condition were asked to do their best to recall 10 pieces of information in an open-ended manner by typing their answer into a text box, with up to 25 seconds to answer each. Queried items included such things as the name of the man in the story, and what he threw on the laptop (see the Supplemental Material<sup>16</sup>).

Next, all participants were presented with one of two screens displaying new information about Francis. In the control condition, participants read information consistent with their prior negative impression: Francis had a criminal record, often screamed at kids playing near his property, and broke into the houses in search of valuables and revenge on the children. In the fire

---

<sup>16</sup> Provided in an attached Appendix.

rescue story condition, participants instead read information that provided a highly positive reframing of his earlier actions: Francis broke into the homes because he saw that they were on fire, and the “precious things” he removed were the young children who were trapped inside. Participants were asked to think about this new information for at least fifteen seconds before advancing.

Participants then completed the AMP again, identical to the first (except using the remaining ideograph set). Next, they recompleted the explicit attitude scale, and reported whether they spoke Mandarin or Cantonese. The remaining questions were exploratory, and thus not analyzed here. They answered three multiple-choice questions assessing their final interpretations of events in the story, including why Francis threw water (e.g., to put out a fire; to ruin items), why the cat died (e.g., smoke inhalation; injuries from getting stepped on), and what Francis removed from the houses (e.g., children; jewelry). Participants were then asked to identify Francis out of a lineup of the photographs of men presented during the study, one of which was Francis West, and the other 10 of which were the control primes. After this, they were asked how confused they were (from 1, not confused at all, to 7, completely confused), whether they thought the story was based on true events (from 1, not at all, to 7, completely), and how they feel right now (from 1, very bad, to 7, very good). They were also given a manipulation check asking them to identify the final information they had read about Francis, from a set of three options, and then finally reported demographic information, were offered the chance to provide open-ended feedback to the researchers about the study, and were debriefed and compensated.

## **Results**

All but four participants correctly identified the final information they had read about Francis West on the manipulation check, demonstrating high levels of attention to the critical Time 2 information. All 270 participants were included in the analyses below, but the results do not meaningfully differ if those four participants are excluded.

### **Implicit Evaluations**

Implicit evaluations of the primes (Francis West or control faces) were measured as the proportion of pictographs judged to be more pleasant than average within each cell of a 2 (Measurement Time: Time 1, Time 2) x 2 (Prime Person: Francis West, Control Faces) x 2 (Story Condition: Fire Rescue, Control) x 2 (Recall Condition: Quiz Present, Quiz Absent) mixed design, with the first two factors manipulated within-participants.

A mixed-ANOVA revealed that the highest-order significant effect was the interaction between time, prime person, and story condition,  $F(1,266) = 58.77, p < .001, \eta_p^2 = .181$ . Recall condition did not moderate this effect,  $F(1,266) = .15, p = .703, \eta_p^2 = .001$ , and produced no significant main effect ( $F[1,266] = .388, p = .534, \eta_p^2 = .001$ ) or other interactions, including two-way interactions with time ( $F[1,266] = .335, p = .563, \eta_p^2 = .001$ ), prime person ( $F[1,266] = .530, p = .467, \eta_p^2 = .002$ ), or story condition ( $F[1,266] = .606, p = .437, \eta_p^2 = .002$ ) and three-way interactions with time and prime person ( $F[1,266] = 1.893, p = .170, \eta_p^2 = .007$ ), time and story condition ( $F[1,266] = .216, p = .643, \eta_p^2 = .001$ ), or prime person and story condition ( $F[1,266] = .338, p = .561, \eta_p^2 = .001$ ). Figure 12 shows the mean implicit positivity of each prime type across conditions. Examining the three-way effect, we found that the interaction between time and prime person was significant in the fire rescue condition,  $F(1,266) = 76.56, p < .001, \eta_p^2 = .223$ , such that at Time 1, implicit evaluations of Francis West ( $M = .41, SD = .28$ ) were significantly less positive than of the control faces ( $M = .60, SD = .22$ ),  $F(1,266) = 32.16, p <$



.001,  $\eta_p^2 = .108$ , but at Time 2, implicit evaluations of Francis West ( $M = .65$ ,  $SD = .26$ ) were significantly more positive than of the control faces ( $M = .50$ ,  $SD = .21$ ),  $F(1,266) = 21.11$ ,  $p < .001$ ,  $\eta_p^2 = .074$ . This suggests that reversal of the implicit first impression occurred, and just as strongly as in previous work – the Cohen's  $d$  comparing relative implicit preference for Francis over control faces (difference score) at Time 2 in the fire condition versus the control condition was 1.23, which is comparable to the studies in Mann and Ferguson (2015) that used the same measure (range 1.03 – 1.15). In the control story condition, there was a marginal interaction between time and prime person,  $F(1,266) = 3.70$ ,  $p = .056$ ,  $\eta_p^2 = .014$ . Implicit evaluations of Francis ( $M = .37$ ,  $SD = .29$ ) were significantly less positive than of the control faces ( $M = .61$ ,  $SD = .23$ ) at Time 1,  $F(1,266) = 53.61$ ,  $p < .001$ ,  $\eta_p^2 = .168$ , and this difference was even more pronounced at Time 2 (Francis West:  $M = .37$ ,  $SD = .26$ ; Control:  $M = .68$ ,  $SD = .21$ ),  $F(1,266) = 95.96$ ,  $p < .001$ ,  $\eta_p^2 = .265$ .

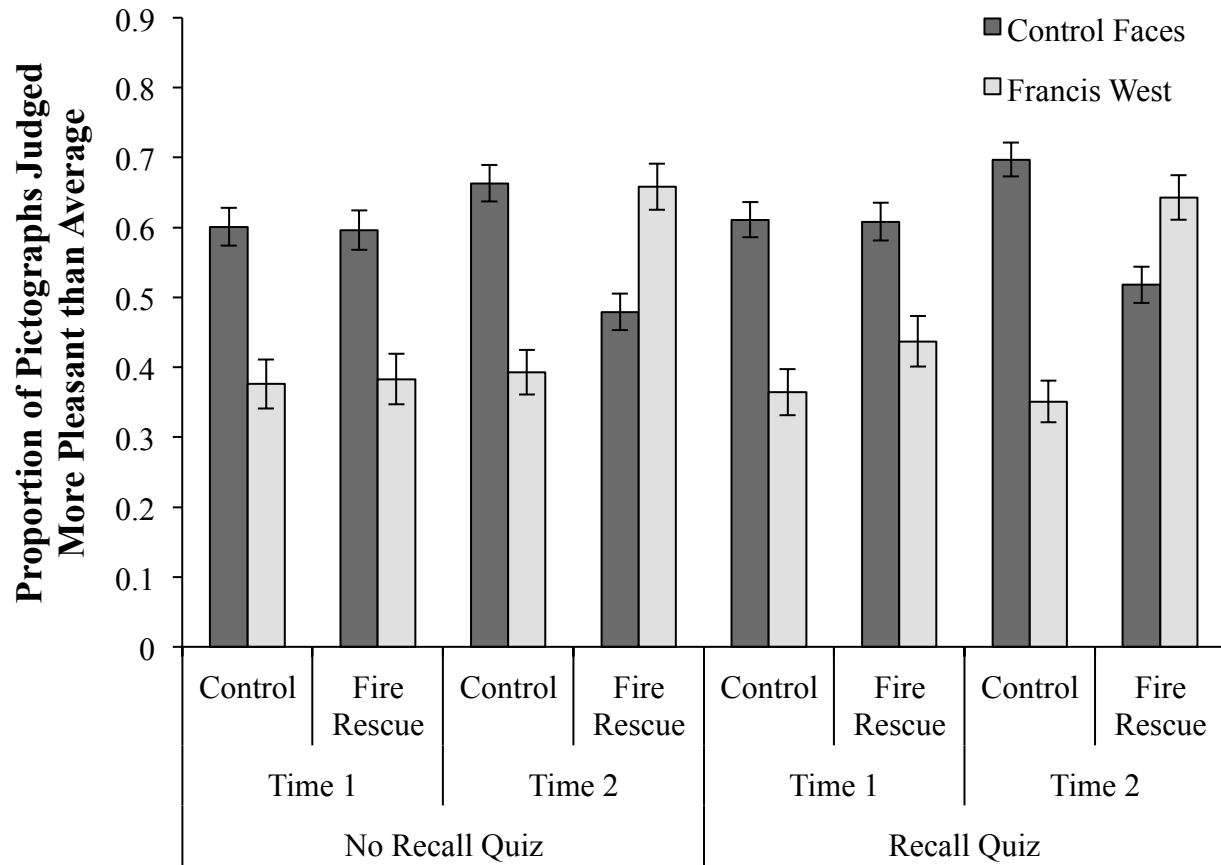


Figure 12. Mean proportion of pictographs judged to be more pleasant than the average pictograph, by quiz condition, time, story condition, and person prime, in Study 9. Error bars are standard errors.

We then tested whether individual performance on the recall quiz predicted revision. Two blind coders separately made yes/no determinations for each question as to whether the participant had correctly recalled the information from session one (98% agreement). We followed an a priori plan to score answers as “correct” only if both coders judged it as correct. For each participant, we computed recall performance as the total number of correct responses ( $M_{\text{Recall}} = 5.18$  of 10,  $SD = 1.84$ , range 0 – 9).

Next, we ran a mixed ANOVA on implicit pleasantness in the recall quiz condition, with all interactions and main effects involving time, person prime, story condition, and the continuous measure of recall performance (mean-centered). Results showed that recall performance did not moderate the three-way interaction between time, person prime, and story condition,  $F(1,136) = .689, p = .408, \eta_p^2 = .005$ , nor was there a significant main effect of performance ( $F[1,136] = 2.056, p = .154, \eta_p^2 = .015$ ) or other interactions, including two-way interactions with time ( $F[1,136] = 0.595, p = .442, \eta_p^2 = .002$ ), prime person ( $F[1,136] = .502, p = .480, \eta_p^2 = .004$ ), or story condition ( $F[1,136] = 0.008, p = .929, \eta_p^2 < .001$ ) and three-way interactions with time and prime person ( $F[1,136] = 0.077, p = .782, \eta_p^2 = .001$ ), time and story condition ( $F[1,136] = .971, p = .326, \eta_p^2 = .007$ ), or prime person and story condition ( $F[1,136] = .844, p = .360, \eta_p^2 = .006$ ).

### **Explicit Evaluations**

Explicit attitudes toward Francis also showed revision in the fire rescue condition. An interaction effect on the mean of the six-item explicit liking scale emerged between time and story condition,  $F(1,266) = 1084.50, p < .001, \eta_p^2 = .80$ , such that Francis was viewed equivalently at Time 1 in the fire condition ( $M = 1.16, SD = .45$ ) and in the control condition ( $M = 1.28, SD = .70$ ),  $F(1,266) = 2.58, p = .109, \eta_p^2 = .01$ , but at Time 2 was viewed more positively in the fire ( $M = 5.81, SD = 1.47$ ) versus control condition ( $M = 1.21, SD = .45$ ),  $F(1,266) = 1251.40, p < .001, \eta_p^2 = .82$ . The interaction between time and story condition was not moderated by recall quiz condition,  $F(1,266) = .260, p = .611, \eta_p^2 = .001$ , and there was also no main effect ( $F[1,266] = .173, p = .678, \eta_p^2 = .001$ ), interaction with time ( $F[1,266] = .691, p = .406, \eta_p^2 = .003$ ), or interaction with story condition ( $F[1,266] = .855, p = .356, \eta_p^2 = .003$ ).

To check for effects of recall performance on explicit evaluations, we conducted a mixed ANOVA in the quiz condition that included all main effects and interactions involving time, story condition, and recall performance. Results showed a marginally significant three-way interaction,  $F(1,136) = 3.168, p = .077, \eta_p^2 = .023$ , such that there was an interaction between recall performance and story condition at Time 2,  $F(1,136) = 4.943, p = .028, \eta_p^2 = .035$ , but (unsurprisingly) not at Time 1,  $F(1,136) = .267, p = .606, \eta_p^2 = .002$ . In the fire condition, greater recall predicted more positive explicit evaluations of Francis at Time 2,  $B = .172, SE = .079, p = .031, \eta_p^2 = .034$ , but there was no relationship between recall and explicit evaluations in the control condition at Time 2,  $B = -.065, SE = .072, p = .367, \eta_p^2 = .006$ .

## **Discussion**

These results demonstrate that a route through which negative implicit first impressions have been reversed shortly after formation – reinterpretation (Mann & Ferguson, 2015) – remains effective even after a delay, despite the potential challenges to such revision through forgetting, trait abstraction, and/or memory consolidation. Participants learned new information about a person two days after forming a first impression, which is longer than all studies to date examining whether implicit evaluations of novel targets can be updated and exceeds the delays used in many studies attempting to edit consolidated memories (e.g., Agren et al., 2012; Schiller et al., 2010). Despite this delay, those given reinterpretation information showed a robust reversal of their implicit impression, including those who were not prompted to recall earlier details as well as those who had poor recall. In the control condition, as predicted, implicit evaluations of Francis remained negative over time.

The results show that implicit evaluations track relevant experiences over time, rather than remain stuck in initial experiences. The results also suggest that implicit evaluations do not

rely on (only) slow-changing associations (e.g., Rydell & McConnell, 2006), and that propositional reasoning can impact implicit processes (e.g., De Houwer, 2014; Gawronski & Bodenhausen, 2006). They are consistent with work on the malleability of implicit evaluations across the lifespan (e.g., Castelli et al., 2010), using experimentally controlled first impressions of novel others to distinguish new learning from reactivation of preexisting contextual attitudes (see Fazio, 2007; Gregg et al., 2006; Cone & Ferguson, 2015; Mann & Ferguson, 2015). This suggests that the window for effective reversal of implicit evaluations does not close immediately (cf. Peters & Gawronski, 2011), at least within a reinterpretation paradigm and with highly relevant new information. Though we can only speculate on the reason for the differences across paradigms, it may be that the form of revision attempted in previous work – simply *negating* the earlier information – may be less effective without replacing the impression with something new. The rejection of the earlier impression may also be perceived as more valid and diagnostic when such a replacement is available, and reinterpretation may constitute an effective case of this combined “subtraction + addition” approach (see discussion in Mann & Ferguson, 2015). Different types of new information may have unique temporal windows during which revision is possible.

Diagnosticity is an important factor driving explicit (Skowronski & Carlston, 1989), as well as implicit impressions (Cone & Ferguson, 2015; Mann & Ferguson, 2015). In our view, new information that reinterprets earlier information can be construed as diagnostic in that it changes what earlier behaviors diagnose about the person (Mann, Cone, & Ferguson, 2015); as such, the current work builds on the importance of diagnosticity, demonstrating its importance not just for implicit impression formation, but also for updating over time. The strength of this revision after two days, regardless of recall cues or performance, suggests that whatever initial

memory consolidation, trait abstraction, and forgetting of specific details about the source of an impression may occur over the course of a couple days do not preclude the possibility of effective implicit impression change through reinterpretation.

Although the current paradigm used new information that completely explained the initial information, it is also possible that new information would only partially do so, and it may be that explicit recall of the initial details determines extent of updating in such cases. For instance, with more ambiguous new information, explicit recall of the initial details might determine extent of updating (see Wyer 2010, 2016). Collectively, this might suggest that *some* reactivation of memories for earlier details is important for revision – in line with work on retrieval-driven memory updating (e.g., Lane et al., 2015) – but that new information that strongly reframes earlier details can intrinsically produce sufficient reactivation without other external prompting. What is clear, however, is that amount of recall *before* learning the new information did not impact the extent of revision in this paradigm, suggesting that reinterpretation may be an effective way to update initial impressions even when some forgetting has taken place. It could be that when new, reinterpreting details are provided after a longer delay, the greater abstraction of impressions would make retrieval of the specific behavioral memories required for reinterpretation less likely (Klein et al., 1992; Sherman & Klein, 1994; see also Hastie & Park, 1986). Additionally, we obtained a (marginal) interaction involving recall performance on explicit evaluations, and so explicit (vs. implicit) evaluations may be more sensitive to recall, though more research on this is needed. It is also possible that revision through other types of information after a delay (besides details that reframe the earlier impression) is more dependent on recall, which future research might explore.

The role of reactivation of earlier learning for revision of various types of responses is a topic of ongoing discussion (Mann, Cone, & Ferguson, 2015; Lane et al., 2015). Future research can more fully test the range of the temporal window during which implicit impression updating is possible, as well as identify cases in which memory recall may moderate revision. For now, our findings add to the emerging literature on when and how implicit first impressions can be updated.

## References

- Agren, T., Engman, J., Frick, A., Bjorkstrand, J., Larsson, E. M., Furmark, T., & Fredrikson, M. (2012). Disruption of reconsolidation erases a fear memory trace in the human amygdala. *Science*, 337(6101), 1550–1552.  
<http://doi.org/10.1126/science.1223006>
- Babey, S. H., Queller, S., & Klein, S. B. (1998). The role of expectancy violating behaviors in the representation of trait knowledge: A summary-plus-exception model of social memory. *Social Cognition*, 16(3), 287–339.
- Bouton, M. E. (1994). Context, ambiguity, and classical conditioning. *Current Directions in Psychological Science*, 3(2), 49–53. <http://doi.org/10.1111/1467-8721.ep10769943>
- Castelli, L., Carraro, L., Gawronski, B., & Gava, K. (2010). On the determinants of implicit evaluations: When the present weighs more than the past. *Journal of Experimental Social Psychology*, 46(1), 186–191.  
<http://doi.org/10.1016/j.jesp.2009.10.006>
- Castelli, L., Zogmaister, C., Smith, E. R., & Arcuri, L. (2004). On the automatic evaluation of social exemplars. *Journal of Personality and Social Psychology*, 86(3), 373.
- Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, 108, 37-57.  
<http://doi.org/10.1037/pspa0000014>
- De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, 8(7), 342–353. <http://doi.org/10.1111/spc3.12111>
- Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled



- components. *Journal of Personality and Social Psychology*, 56(1), 5–18.  
<http://doi.org/10.1037/0022-3514.56.1.5>
- Dudai, Y. (2004). The neurobiology of consolidations, or, how stable is the engram?  
*Annual Review of Psychology*, 55(1), 51–86.  
<http://doi.org/10.1146/annurev.psych.55.090902.142050>
- Dunham, Y., Baron, A. S., & Banaji, M. R. (2008). The development of implicit  
intergroup cognition. *Trends in Cognitive Sciences*, 12, 248–253.
- Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength.  
*Social Cognition*, 25, 603–637. <http://dx.doi.org/10.1521/soco.2007.25.5.603>
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in  
evaluation: An integrative review of implicit and explicit attitude change. *Psychological  
Bulletin*, 132(5), 692–731. <http://doi.org/10.1037/0033-2909.132.5.692>
- Gawronski, B., & Cesario, J. (2013). Of mice and men: What animal research can tell us  
about context effects on automatic responses in humans. *Personality and Social  
Psychology Review*, 17(2), 187–215. <http://doi.org/10.1177/1088868313480096>
- Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in  
the malleability of implicit preferences. *Journal of Personality and Social Psychology*,  
90(1), 1–20. <http://doi.org/10.1037/0022-3514.90.1.1>
- Hastie, R., & Kumar, P. A. (1979). Person memory: Personality traits as organizing  
principles in memory for behaviors. *Journal of Personality and Social Psychology*, 37,  
25–38.
- Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends

- on whether the judgment task is memory-based or on-line. *Psychological Review*, 93(3), 258.
- Kawakami, K., Phills, C. E., Steele, J. R., & Dovidio, J. F. (2007). (Close) distance makes the heart grow fonder: Improving implicit racial attitudes and interracial interactions through approach behaviors. *Journal of Personality and Social Psychology*, 92(6), 957–971. <http://doi.org/10.1037/0022-3514.92.6.957>
- Klein, S. B., Loftus, J., Trafton, J. G., & Fuhrman, R. W. (1992). Use of exemplars and abstractions in trait judgments: A model of trait knowledge about the self and others. *Journal of Personality and Social Psychology*, 63(5), 739.
- Kumkale, G. T., & Albarracín, D. (2004). The sleeper effect in persuasion: A meta-analytic review. *Psychological Bulletin*, 130(1), 143–172. <http://doi.org/10.1037/0033-2909.130.1.143>
- Lane, R. D., Ryan, L., & Nadel, L. (2015). Memory reconsolidation, emotional arousal and the process of change in psychotherapy: New insights from brain science. *Behavioral and Brain Sciences*, 38, 1–64.
- Mann, T., Cone, J., & Ferguson, M. J. (2015). Social-psychological evidence for the effective updating of implicit attitudes. *Behavioral and Brain Sciences*, 38, 32-33.
- Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions? The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology*, 108(6), 823–849. <http://doi.org/10.1037/pspa0000021>
- McCarthy, R. J., & Skowronski, J. J. (2011). What will Phil do next? *Journal of Experimental Social Psychology*, 47(2), 321–332.
- McConnell, A. R., & Rydell, R. J. (2014). The systems of evaluation model: A dual-

- systems approach to attitudes. In J. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual process theories of the social mind* (pp. 204-217). New York: Guilford.
- McGaugh, J. L. (2000). Memory - a century of consolidation. *Science*, 287(5451), 248–251. <http://doi.org/10.1126/science.287.5451.248>
- Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, 36, 630–633.
- Payne, B. K., Brown-Iannuzzi, J., Burkley, M., Arbuckle, N. L., Cooley, E., Cameron, C. D., & Lundberg, K. B. (2013). Intention Invention and the Affect Misattribution Procedure: Reply to Bar-Anan and Nosek (2012). *Personality and Social Psychology Bulletin*, 39(3), 375–386. <http://doi.org/10.1177/0146167212475225>
- Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89(3), 277–293. <http://doi.org/10.1037/0022-3514.89.3.277>
- Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin*, 37(4), 557–569. <http://doi.org/10.1177/0146167211400423>
- Pratkanis, A. R., Greenwald, A. G., Leippe, M. R., & Baumgardner, M. H. (1988). In search of reliable persuasion effects: III. The sleeper effect is dead: Long live the sleeper effect. *Journal of Personality and Social Psychology*, 54, 203–218.
- Rudman, L. A., Phelan, J. E., & Heppen, J. B. (2007). Developmental sources of implicit attitudes. *Personality and Social Psychology Bulletin*, 33, 1700–1713.
- Rule, N. O., Ambady, N., Adams, R. B., Jr., Ozono, H., Nakashima, S., Yoshikawa, S., &

- Watabe, M. (2010). Polling the face: Prediction and consensus across cultures. *Journal of Personality and Social Psychology*, 98, 1-15.
- Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit attitude change: A systems of reasoning analysis. *Journal of Personality and Social Psychology*, 91(6), 995–1008. <http://doi.org/10.1037/0022-3514.91.6.995>
- Schiller, D., Kanen, J. W., LeDoux, J. E., Monfils, M. H., & Phelps, E. A. (2013). Extinction during reconsolidation of threat memory diminishes prefrontal cortex involvement. *Proceedings of the National Academy of Sciences*, 110(50), 20040–20045. <http://doi.org/10.1073/pnas.1320322110>
- Schiller, D., Monfils, M. H., Raio, C. M., Johnson, D. C., LeDoux, J. E., & Phelps, E. A. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature*, 463(7277), 49–53.
- Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, 105(1), 131-142.
- Sherman, J. W., & Klein, S. B. (1994). Development and representation of personality impressions. *Journal of Personality and Social Psychology*, 67(6), 972-983.
- Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: Evidence from a false recognition paradigm. *Journal of Personality and Social Psychology*, 83(5), 1051–1065. <http://doi.org/10.1037//0022-3514.83.5.1051>
- Van Dessel, P., De Houwer, J., Gast, A., & Smith, C. T. (2015). Instruction-based

- approach–avoidance effects: Changing stimulus evaluation via the mere instruction to approach or avoid stimuli. *Experimental Psychology*, 62, 161–169.  
<http://dx.doi.org/10.1027/1618-3169/a000282>
- Van Dessel, P., De Houwer, J., Gast, A., Smith, C. T., & De Schryver, M. (2016). Instructing implicit processes: When instructions to approach or avoid influence implicit but not explicit evaluation. *Journal of Experimental Social Psychology*, 63(C), 1–9.  
<http://doi.org/10.1016/j.jesp.2015.11.002>
- Westfall, J. (2015). PANGEA: Power ANalysis for GEneral Anova designs (Working paper). Retrieved from: <http://jakewestfall.org/publications/pangea.pdf>
- Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, 107(1), 101.
- Woud, M. L., Maas, J., Becker, E. S., & Rinck, M. (2013). Make the manikin move: Symbolic approach–avoidance responses affect implicit and explicit face evaluations. *Journal of Cognitive Psychology*, 25(6), 738–744.  
<http://doi.org/10.1080/20445911.2013.817413>
- Wyer, N. A. (2010). You never get a second chance to make a first (implicit) impression: The role of elaboration in the formation and revision of implicit impressions. *Social Cognition*, 28(1), 1–19.
- Wyer, N. A. (2016). Easier done than undone... by some of the people, some of the time: The role of elaboration in explicit and implicit group preferences. *Journal of Experimental Social Psychology*, 63, 77–85.
- Zanon, R., De Houwer, J., Gast, A., & Smith, C. T. (2014). When does relational

information influence evaluative conditioning? *The Quarterly Journal of Experimental Psychology*, 67(11), 2105–2122. <http://doi.org/10.1080/17470218.2014.907324>

## **Appendix: Text of Time 2 Recall Quiz Questions, Study 9**

### **Instructions**

You may recall from a few days ago that you read details from a story. We are going to ask you a series of questions about that story to see how much you remember. You have up to 25 seconds to answer each question. Please just answer from what you remember; don't look anything up.

*Proceed to the next page to begin.*

### **Questions**

1. In the story, what was the name of the man who broke into the houses in his neighborhood?
2. What type of animal did the man step on in one of the houses?
3. How many houses did the man break into?
4. What did he throw onto a laptop in the kitchen?
5. What did he use to break down the door of the first house?
6. What did the man knock over and shatter in the first house?
7. What did the man walk through on his way to the second house?
8. At the second house, what did the man use to break through the stained-glass window?
9. What was the name of the pet at the second house?
10. What did the man step on and walk across in the basement of the second house?

### **Study 10: Revision with Real-World Targets**

The paper presented as Study 9 examined one form of the broader generalizability and utility of reinterpretation as a route for updating implicit evaluations, finding that such revision is still efficacious days after the formation of an initial impression of Francis West. Together with the result of Study 6 in Chapter II, which found that revised evaluations endured for days after revision, this work suggests that the passing of time need not be an uncompromisingly limiting factor in the potential for implicit change to occur and endure. Given that information about many real social entities also flows in over time (rather than a single lab session) and that impressions formed of such targets can also impact behavior over a broader span of time, these findings were initial tests of a critical element of the general applicability of reinterpretation-driven change.

In Study 10, the external validity of the basic reinterpretation effects found in Chapter II is tested in a very different manner, by examining how initially negative implicit impressions of a real person are affected by arguments that offer a reinterpretation of the actions of that person. Importantly, the actions of this person, and the arguments offered that attempt to reinterpret those actions, were drawn from true, external events. This allows a test of the ability of reinterpretation to update implicit impressions in a realistic paradigm, in which the actions under discussion were naturally produced and the arguments made about them were genuinely designed to sway public opinion.

Specifically, the study introduces participants to a big-game hunter who attracted public controversy when he won an auction for the right to go abroad to hunt and kill an endangered black rhino, an act that he subsequently completed (Lavandera, 2015). To the degree that participants view that activity negatively, this is expected to form an initially negative implicit



evaluation of the hunter. The study then presents participants with a clip from the podcast Radiolab (<http://www.radiolab.org/>), which either offers a positive reinterpretation of this practice as beneficial to conservation by helping countries raise money to fund it, or describes an unrelated topic in a control condition. The goal of the study is to then assess how effectively this argument updates implicit evaluations, and whether any change that is observed is mediated by self-reported reinterpretation.

Importantly, because the podcast offers a possible new interpretation of big-game hunting that – in comparison to the reinterpretation of the actions of Francis West – is less compulsory to endorse (given the possibility of counter-arguments), more motivationally and ideologically charged, and only one element of the behavior in question (because of the possibility that the actions of the hunter are driven not just by a passion for conservation but also for other reasons, like intrinsic enjoyment of the activity itself), the study provides an opportunity to test whether reinterpretation is still effective in shifting implicit evaluations in a “messier” real-world context. Due to this difference from the Francis West paradigm, a wholesale reversal of implicit evaluations of the hunter after participants listen to the podcast is not expected, but a significant shift is. For some (or many) participants, the new information might not prompt reinterpretation (or have a negate + affirm effect) at all, and at most will add one new reason to view hunting less negatively while not negating all of their reasons for disliking the practice of big-game hunting or hunters themselves; for that reason, the new information could be most similar to the subway rescue condition in the Francis West work, rather than the fire rescue (reinterpretation) condition. To the extent that reinterpretation does occur after the podcast, however, the effect of podcast on implicit impressions will be at least partially mediated by self-reported reinterpretation, conceptually replicating the result of Study 5 in a very different domain.

## Method

**Participants.** I planned to recruit as many participants as we could for this study during a single semester, resulting in a sample of 213 Cornell University students who participated in exchange for partial course credit. Of these, 3 were excluded for failing to select the correct summary of the podcast they listened to during the study, and 7 more were excluded for using a single key on every trial of at least one of the two AMPs (both exclusion criteria were determined a priori), resulting in a final sample size of 203 (76% women, Age  $M = 19.99$  years,  $SD = 1.33$ ). Participants completed the experiment in individual computer rooms.

**Initial learning.** At the beginning of the study, participants viewed a single screen of information about Corey Knowlton in order to familiarize them with his name, face, and activities surrounding big-game hunting that were covered in the media. In particular, they were informed that Knowlton had paid \$350,000 for a permit to hunt and kill a black rhino in Namibia, and that in part due to the status of that species as endangered, this action had attracted controversy in the media (Lavandera, 2015). Nonetheless, Knowlton completed his hunt in 2015. This initial learning task was designed to foster a negative impression of Knowlton by associating him with an activity (big-game hunting of endangered species) to which most participants would likely have an initially negative reaction, based on the results of a pilot test.

**First AMP.** Immediately following the initial learning task, participants completed an AMP similar to those used in my earlier studies, with two major differences: the composition of the primes, and the nature of the targets. On 20 trials, the prime image consisted of a frontal view of Corey Knowlton's face and upper body, selected from search engine results for its relatively neutral expression, cropped and displayed on a white background. On each of the other 20 trials, the prime image consisted of one of 5 control faces of white males on white

backgrounds (each presented on 4 trials), drawn from a publicly available face database used in prior research (The Chicago Face Database; Ma, Correll, & Wittenbrink, 2015).

The second change related to the targets. In my prior studies using online samples, the target stimuli consisted of Chinese pictographs drawn from prior research with this task (Payne et al., 2005; Payne & Lundberg, 2014), as these stimuli had been demonstrated to be generally close to neutral in visual pleasantness for most participants, a desirable feature for a misattribution task. However, the use of such stimuli requires the exclusion of participants who are familiar with the meaning of the pictographs. On Mechanical Turk, few participants fall into this category, but pilot testing indicated that a number of students in the undergraduate participant pool used in Study 10 would need to be excluded prior to analysis. To avoid losing these participants, I instead used a set of target stimuli that I initially developed with my colleagues for use in a different project (Mann, Katz, Ferguson, & Goncalo, in preparation). These consisted of 80 “paintings” that all featured a solid color background and numerous colorful lines. Like the pictographs, the paintings were made visually similar to minimize the likelihood that participants would have strong inherent preferences between them, so as to increase the relative signal from misattribution of prime evaluations in their AMP responses (see examples in Figure 13). The paintings were randomly assigned to one of two sets, and for each participant, one set was selected for use during this first AMP, with the unselected set employed on the second AMP.



Figure 13. Example painting stimuli from Study 10.

**Explicit evaluations.** Participants next responded to six 7-point Likert-type scale items, asking them to place Corey Knowlton on scales from *unlikable* to *likable*, *bad* to *good*, *mean* to *pleasant*, *disagreeable* to *agreeable*, *uncaring* to *caring*, and *cruel* to *kind*.

**Podcast.** Next, participants were randomly assigned to listen to one of two eight-and-a-half minute podcast excerpts. The excerpt used in the experimental condition was material drawn from the episode “The Rhino Hunter” of the podcast *RadioLab* (2015), which described a view of big-game hunting as having a *positive* impact on the environment and the survival of endangered species under some conditions. In brief, the argument espoused by some hunters, African governments, and conservation organizations is that the preservation of many endangered species requires levels of funding that have proven difficult to maintain in the face of

poachers, budgetary requirements for paying game wardens and buying patrol vehicles and necessary equipment, and the unwillingness of communities to tolerate living in close proximity to dangerous animals. Governments auctioning off the right to hunt and shoot individual animals (which are often chosen due to being old, post-reproductive, and a danger to other members of the species) creates an economic value for the survival of the species and provides a source of revenue for conservation efforts. The full podcast episode includes extended interviews with Corey Knowlton himself, details of his hunt, as well as some discussion from critics of these arguments.

To create a test of whether arguments that provide a reinterpretation of views on a real issue (hunting of endangered species) could shift initially negative implicit impressions of a person associated with that issue, we constructed an 8.5-minute excerpt of this podcast episode, retaining only those segments identified by a team of research assistants as most effectively supporting the view of this form of hunting as pro-conservation. For the control condition, we produced a clip of the first 8.5 minutes of a different episode of *RadioLab*, describing the origins of units of measurement (“Weights and Measures”).

**Second AMP and explicit evaluations.** After listening to either the experimental or control podcast, participants completed the AMP again, identical to the first, followed by the same explicit evaluation measures used prior to the podcast. They were also asked to identify the correct summary of the podcast they heard from a set of 4 options.

**Reinterpretation questions.** Participants answered three questions designed to measure the extent to which the podcast they listened to prompted them to reinterpret their views on big-game hunting. These included, “When you listened to the podcast clip, how much did the information in that clip change the meaning of Corey Knowlton's actions that you learned about

at the beginning of the study?” “To what degree did the podcast clip that you listened to earlier in this study change how you think about big-game hunting?” and “After you listened to the podcast clip earlier in this study, to what extent did you see big-game hunting in a different light?” all on scales from 1 (*Not at all*) to 9 (*Completely*).

**Exploratory items.** Finally, participants responded to two questions included for exploratory purposes, which for the sake of space will not be discussed further:

“How convinced were you by the arguments that were made in the podcast that you listened to earlier in this study?” on a scale from 1 (*Not at all convinced*) to 9 (*Extremely convinced*), and “When you were listening to the podcast clip earlier in this study, how often were you thinking about arguments against the views expressed in the podcast?” on a scale from 1 (*Not at all*) to 9 (*The entire time*).

## Results

**Implicit evaluations.** Responses on the AMPs were assessed in a 2 (Time: Time 1, Time 2) x 2 (Prime: Corey Knowlton, Control Faces) x 2 (Podcast: Hunting, Control) mixed ANOVA, with the first two factors manipulated within-subjects. Scores within each cell represent the proportion of paintings judged to be more pleasant than average following the prime.

A significant three-way interaction between time, prime, and podcast obtained,  $F(1, 201) = 7.025, p = .009$ . At Time 1, Knowlton was less implicitly positive than control faces in both the hunting podcast condition (Knowlton:  $M=.39, SD=.20$ ; control faces:  $M=.55, SD=.19$ ),  $t(100) = 5.67, p < .001$ , and the control podcast condition (Knowlton:  $M=.40, SD=.19$ ; control faces:  $M=.57, SD=.19$ ),  $t(101) = 6.52, p < .001$ . At Time 2, in the control podcast condition, Knowlton was less implicitly positive than control faces (Knowlton:  $M=.40, SD=.22$ ; control faces:  $M=.56, SD=.22$ ),  $t(101) = 5.46, p < .001$ . However, in the hunting podcast condition, though Knowlton

was still less implicitly positive than control faces, the difference was much smaller and closer to neutral (Knowlton:  $M = .47$ ,  $SD = .22$ ; control faces:  $M = .52$ ,  $SD = .22$ ),  $t(100) = 2.09$ ,  $p = .039$ .

Figure 14 shows average implicit evaluations within each cell of the overall design.

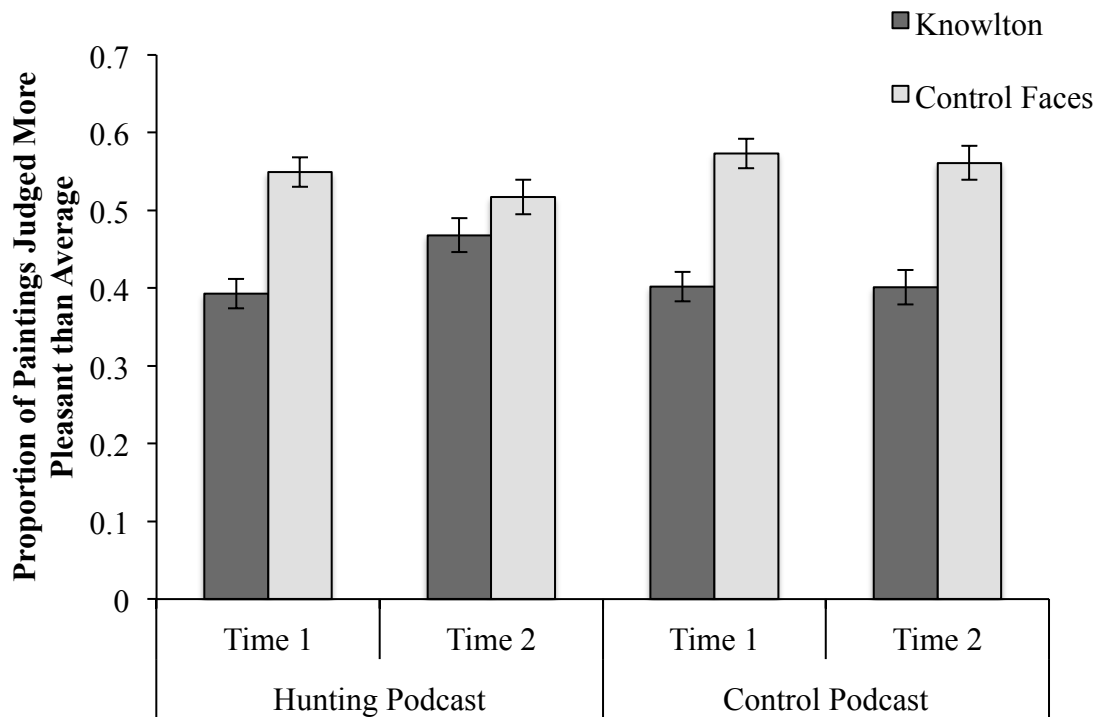


Figure 14. Mean proportion of paintings judged more pleasant than average, by podcast condition, time, and prime type, Study 10. Error bars are standard errors.

**Explicit evaluations.** Explicit evaluations of Knowlton were analyzed in a 2 (Time: Time 1, Time 2) x 2 (Podcast: Hunting, Control) mixed ANOVA. There was a significant interaction between time and podcast,  $F(1,201) = 213.18$ ,  $p < .001$ . Explicit evaluations of Knowlton were significantly below the midpoint of the scale (4) in all conditions, but much more so at Time 1 and in the control podcast condition at Time 2 ( $ps < .001$ ) than in the hunting podcast condition at Time 2 ( $p = .034$ ).

**Mediation by reinterpretation.** The three questions assessing the extent to which the podcast prompted reinterpretation of views on big-game hunting partly can be interpreted as a manipulation check, as levels of reinterpretation must almost certainly be lower in the control podcast condition. However, they can be additionally assessed as mediators of the podcast effect on implicit change. The three measures correlated highly (Cronbach's  $\alpha = .97$ ), even within podcast conditions (Alphas of .81 and .87), so I merged them into one index score of reinterpretation.

In a mediation model looking at podcast condition effects on the double difference score reflecting change in relative implicit liking of Knowlton (vs. control faces) from Time 1 to Time 2, mediated via the reinterpretation item, the mediation effect was significant, 95% CI: [.0283, .2690], Sobel  $p = .025$ . The inclusion of the mediator reduces the condition effect to non-significance,  $p = .523$ , while the mediator remains significant,  $p = .025$ . Thus, the pattern is consistent with full mediation of the podcast effect on shifts in implicit evaluations of Corey Knowlton through reinterpretation.

I should note that the reinterpretation questions were asked after the final AMP, and the reverse mediation model (condition  $\rightarrow$  AMP double difference score  $\rightarrow$  reinterpretation index score) was also supported, 95% CI: [.0126, .1942], though the Sobel test was marginally significant ( $p = .09$ ) and the mediator did not eliminate the condition effect in this model ( $p < .001$ ), such that this alternative direction of the effect would only be consistent with partial mediation.

## **Discussion**

This study examined the potential for reinterpretation to produce revision in implicit evaluations in a more naturalistic context than the Francis West paradigm, and using arguments



that were specifically developed by an external party to inform and persuade. The results showed, unsurprisingly, that revision was much milder than in the Francis West story; there are countless reasons why this may be, which the numerous differences between this paradigm and the Francis West studies do not allow me to isolate. Critically, however, a shift in implicit evaluations of Knowlton did occur after the presentation of the arguments in the hunting podcast. This supports the idea that the arguments presented about a negative topic are capable of impacting implicit impressions; furthermore, the mediation pattern involving self-reported reinterpretation was consistent with reinterpretation being the operative mechanism through which the effects of the podcast on implicit evaluations were produced.

A primary goal of future work in this line should be to determine the factors that moderate whether the new information about big-game hunting (or other topics) can actually reverse implicit impressions to a degree present in the Francis West paradigm. Although it is possible that new arguments about real-world issues can never be as broadly effective as the revelations in the Francis West paradigm, perhaps due to ideological convictions or the continued relevance of other factors to the impressions under study, it seems more likely that such factors (when identified) may provide insight into the conditions under which robust change may be possible across a broader sample—for instance, if arguments can avoid ideological statements, appeal to both liberals and conservatives, address lingering concerns, etc. The mediation analysis in the current work, identifying reinterpretation as a conduit for revision toward Knowlton through the arguments of the podcast, provides a plausible mechanism through which such moderators may impact the size of revision.

In the next study, the goal to examine the broader generalizability of implicit revision through reinterpretation returns to the Francis West paradigm, by testing whether reinterpreting

information about Francis continues to be effective when the initial, negative impression aligns with group-based stereotypes.

### **Study 11: Counter-Stereotypical Impressions – I**

The studies in Chapter II were designed to establish the mere *possibility* of the revision of a negative implicit first impression, and pinpoint reinterpretation as the mechanism for such updating; for this purpose, the Francis West paradigm was constructed specifically such that the new information learned about the character of Francis West would completely reframe the entire source of initial negativity toward him – the breaking-and-entering story. The results of Study 10, in showing far weaker (albeit still significant) revision in a context in which the reinterpretation is less wholesale and clear-cut, raise the more general question of how delicate the revision effect is in the face of complicating factors. Where Study 10 examined the implications of added complexity in the form of less comprehensive or less universally convincing reinterpretation, the present study examines complexity in the initial formation of the first impression itself. Specifically, this study addresses the question of whether the implicit first impression of Francis West can be reversed when he is depicted as Black, and therefore a member of a group that is implicitly stereotyped and evaluated in a manner consistent with the original, to-be-overturned impression (i.e. stereotypically hostile and untrustworthy, and evaluatively negative; Greenwald & Banaji, 1995; Nosek et al., 2002).

Decades of research in social cognition have found that implicit evaluations in America remain markedly pro-White and anti-Black, despite sweeping declines in explicit prejudice and stereotyping (Banaji & Greenwald, 2013; Nosek et al., 2002). Americans possess an ordered hierarchy of implicit preferences between social groups on numerous dimensions like age, race, and religion (Axt, Ebersole, & Nosek, 2014); when it comes to race, Blacks are implicitly

evaluated less positively than Whites overall, with such biases observable from a young age (Baron & Banaji, 2006). Such bias prominently manifests in the tendency to perceive anger more readily in Black faces (Hugenberg & Bodenhausen, 2003, see also Hugenberg & Bodenhausen, 2004; Hutchings & Haddock, 2008), and a bias toward more readily perceiving (or misperceiving) weapons when paired with Blacks (Payne, Lambert, & Jacoby, 2002) – a bias which is also revealed in split-second “shoot vs. don’t shoot” decision making in simulations of the decisions police officers must make as to whether a Black vs. White suspect holds a gun vs. a harmless object (Correll, Park, Judd, & Wittenbrink, 2002). Occurrences of such group-based stereotyping and prejudice in implicit cognition thus reflect cultural stereotypes of Blacks as more hostile, aggressive, and generally negative than Whites (Eberhardt, Goff, Purdie, & Davies, 2004; Sagar & Schofield, 1980; Wittenbrink, Judd, & Park, 1997).

Some evidence suggests that group-based implicit impressions such as those derived from race may make it particularly difficult for reinterpretation to update implicit impressions in a counter-stereotypic way. McConnell and colleagues (2008) found that although implicit impressions reflected the evaluative implications of the behaviors of individuals when the images of those individuals showed them to be White, of normal weight, and average attractiveness, implicit evaluations of Black, overweight, or particularly attractive/unattractive faces were impacted only by those visual cues rather than individuating information. The researchers argued that although implicit evaluations could accommodate propositional knowledge in the absence of evaluatively charged “associative” visual cues, those cues have a privileged degree of access to implicit processes when they *are* available, although the exact demarcation of what makes a feature associative vs. nonassociative, and the parameters of interaction between the dual systems in their conception remain unclear (Ferguson, Mann, & Wojnowicz, 2014; cf.

McConnell & Rydell, 2014). Nonetheless, the work of McConnell and colleagues (2008) raises the possibility that the heroic reinterpretation of the actions of Francis West may be far less effective in reversing implicit evaluations when visual features of Francis mark him as a member of a group stereotyped in a manner consistent with the initial impression. This could occur if visual or group-based cues have privileged importance in driving implicit responses (McConnell et al., 2008; Rule et al., 2014), or if stereotypic or prejudice-based expectations about his group leads to confirmation bias that disposes participants to form stronger initial negative impressions (Fiske & Neuberg, 1990), though the effects of race on judgment are often largest when information is minimal or ambiguous (e.g., Dovidio & Gaertner, 2000), which the Francis West story is not.

To test whether revision in implicit evaluations through reinterpretation is still possible when the initial implicit impression is compatible with race-based biases, Study 11 manipulates the race of Francis West. Additionally, the study varies the race of the control faces – including both Black and White male faces as controls – to begin to test another element of the general consequences of reinterpretation-driven change: the extent to which implicit updating generalizes to other individuals from the same group as the target; in this case, people of the same race. A general finding in the literature on stereotyping and prejudice is that individual exemplars who violate an expectation about a group are often subtyped or treated as exceptions that do not invalidate the stereotype (Johnston & Hewstone, 1992; Kunda & Oleson, 1995, 1997; Richards & Hewstone, 2001; Weber & Crocker, 1983). Evidence also suggests that generalization is less likely when an individual's actions are relatively extreme in defying a stereotype, as more moderate violations make a group member appear more typical and their actions thus more generalizable (Queller & Smith, 2002). On the other hand, some studies have found *implicit*

generalization of new information about one target to related targets, even in cases in which explicit generalization does not occur; for example, findings have shown that persuasive messages about the color green generalize to implicit evaluations of the brand Heineken (Horcajo, Briñol, & Petty, 2010), evaluations of a brand generalize to a implicit impressions of a new product from that brand even in the presence of negative information about that product (Ratliff, Swinkels, Klerx, & Nosek, 2012), and – most critically for the present discussion – the behavior of one Black person produced implicit evaluations that generalized to unrelated Black people (Ratliff & Nosek, 2011; see also Ranganath & Nosek, 2008). However, that work found some suggestion of stronger generalization of negative information than positive information (a pattern in the literature on explicit generalization as well; Dolderer, Mummendey, & Rothermund, 2009), and the studies informed participants that the new individuals came from the same social groups as the people about whom they had learned, which could have increased generalization. Ultimately, then, it is an open question as to whether a strong reinterpretation that reverses implicit impressions of one individual will generalize to other members of a group. By manipulating the racial group membership of Francis West and including Black and White control faces on the AMP, the present study could begin to test these possibilities.

## **Method**

**Participants.** Three hundred participants from Amazon’s Mechanical Turk completed the study (sample size determined a priori), with 15 excluded for using a single key on every trial of at least one AMP. No participants were excluded for prior knowledge of the pictographs. A further 3 were dropped from all analyses due to a server timeout error that prevented AMP data from being recorded. This left a final sample of 282 (50.7% women; Age  $M = 36.85$  years,  $SD =$

11.95; 85% White, 5.3% Black, 6.0% Asian, 6% Latino, 3.2% other; Percentages add to over 100 because participants could select more than one racial or ethnic identity).

**Procedure.** The study design conformed closely to the basic Francis West procedure used in Chapter II, including the same Francis West story from Studies 1a and 2-6 and an AMP and explicit evaluation scale before (Time 1) and after (Time 2) the final information, with a few modifications. First, the race of main character was manipulated, such that half of participants learned about a Black character during the course of the study, and the other half learned about a White character. To this end, the name of the character was altered from “Francis West” to “Frank West” after feedback from a number of members of the research team suggested that this name sounded more racially neutral. For participants in the White-Frank condition, the image of Frank West was randomly selected (on a per-participant basis) from the same set of 11 faces used in the studies from Chapter II. For participants in the Black-Frank condition, the image was selected from a set of 11 Black male faces from the same database within the same age range as the White face set (20s-early 30s).

The structure of each AMP was adjusted to allow a comparison of the target (Frank West) to control faces of each race. The image of Frank West served as the prime stimulus on each of 30 AMP trials (per administration of the AMP). The remaining 30 trials of the task served as control trials, half with Black primes and half with White primes. On 15 control trials, one of the 10 unselected Black faces was presented (5 randomly selected to be presented twice each, and the other 5 once each); on the other 15, one of the 10 unselected White faces was presented (5 randomly selected to be presented twice each, and the other 5 once each). Two sets of 60 pictographs were used, with one randomly selected for use on the first AMP and the other

for use on the second AMP, per participant. Otherwise, the AMPs were identical to those used in prior studies.

A final departure from earlier studies was the inclusion of an exploratory measure of how guilty participants felt after learning the final information about Frank West, which participants responded to on a Likert-type scale ranging from 1 (*Not at all guilty*) to 7 (*Very guilty*).

Participants were placed either in the original control condition, in which Frank carries out an action consistent with his prior behavior (throwing rocks at the two houses) or the fire rescue condition, in which the purpose behind his actions is revealed to be saving children from a fire.

## Results

**Implicit evaluations.** The proportion of pictographs judged to be more pleasant than average on each AMP was computed for each participant within each of the three prime types. These proportions were then analyzed within a 2 (Time: Time 1, Time 2) x 3 (Prime Type: Frank West, Black Control Faces, White Control Faces) x 2 (Information Condition: Fire Rescue, Control) x 2 (Frank Race: White, Black) mixed ANOVA, with time and prime type manipulated within-participants, and information condition and Frank's race manipulated between-participants.

This analysis revealed the usual three-way interaction between time, prime type, and information condition,  $F(1.70, 472.95) = 45.21, p < .001, \eta_p^2 = .140$  (This and the following analyses used the Huynh-Feldt correction for violation of sphericity). This was qualified by a

four-way interaction with Frank's race,  $F(1.70, 472.95) = 3.96, p = .026, \eta_p^2 = .014$  (see Figure 15).<sup>17</sup>

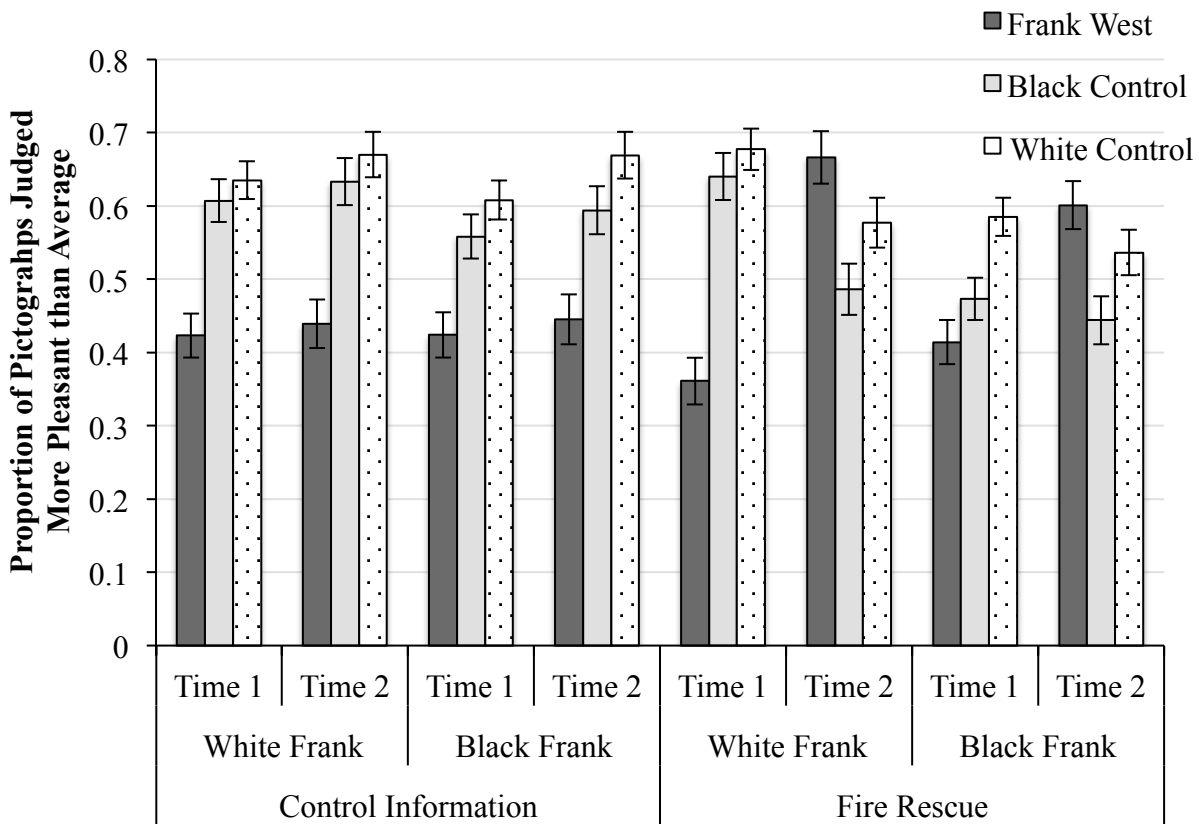


Figure 15. The proportion of pictographs judged to be more pleasant than the average pictograph in Study 11, by information condition, Frank's race, time, and person prime. Error bars are standard errors.

Examining the four-way interaction revealed that the three-way revision effect (between time, prime type, and information condition) was larger when Frank was White,  $F(1.50, 204.30)$

<sup>17</sup> These results, and all others in the present study, did not meaningfully differ when restricting exclusively to White-identified participants, so only analyses using the full sample will be discussed. The three-way interaction between time, prime type, and information condition among White participants was significant,  $F(1.74, 408.21) = 34.08, p < .001, \eta_p^2 = .127$ , as was the four-way moderation by Frank's race,  $F(1.74, 408.21) = 3.75, p = .030, \eta_p^2 = .016$ .



= 30.01,  $p < .001$ ,  $\eta_p^2 = .181$ , than when he was Black,  $F(1.87, 265.26) = 15.48$ ,  $p < .001$ ,  $\eta_p^2 = .098$ , though the effect was clearly significant in both race conditions.

**Initial formation at Time 1.** I first examined the formation of implicit evaluations at Time 1, before the manipulation of final information about Frank West. When Frank was White, he was less implicitly positive ( $M = .39$ ,  $SD = .27$ ) than White control faces ( $M = .65$ ,  $SD = .21$ ),  $t(137) = 8.02$ ,  $p < .001$ , Hedges'  $g_{av} = 1.065$ , and also less implicitly positive than Black control faces ( $M = .62$ ,  $SD = .26$ ),  $t(137) = 6.72$ ,  $p < .001$ , Hedges'  $g_{av} = .852$ . A similar pattern was found when Frank was Black: He was less implicitly positive ( $M = .42$ ,  $SD = .25$ ) than White control faces ( $M = .60$ ,  $SD = .23$ ),  $t(143) = 5.44$ ,  $p < .001$ , Hedges'  $g_{av} = .734$ , and less implicitly positive than Black control faces ( $M = .51$ ,  $SD = .24$ ),  $t(143) = 4.20$ ,  $p < .001$ , Hedges'  $g_{av} = .391$ .

A difference emerged when comparing the Black and White *control faces* based on Frank's race: When Frank was White, there was no difference in implicit responses to Black and White control faces,  $t(137) = 1.53$ ,  $p = .129$ , Hedges'  $g_{av} = .132$ . However, when Frank was Black, White control faces were implicitly more positive than Black control faces,  $t(143) = 2.96$ ,  $p = .004$ , Hedges'  $g_{av} = .344$ . A test of the interaction between these two levels of prime type at Time 1 by Frank's race condition, however, found that this difference between the Black and White control trials based on Frank's race did not reach significance,  $F(1, 280) = 2.06$ ,  $p = .152$ ,  $\eta_p^2 = .007$ .

**Revision at Time 2.** After reading the final information about Frank West, participants in the control story condition continued to have a negative implicit impression of Frank relative to the Black and White control faces for both White and Black Frank versions, all  $ts > 3.8$ , all  $ps < .001$ . Participants in the fire rescue condition, however, showed reversals in their implicit evaluations. When Frank was White, he was marginally more implicitly positive ( $M = .67$ ,  $SD =$

.28) than White control faces ( $M = .58, SD = .29$ ),  $t(62) = 1.72, p = .092$ , Hedges'  $g_{av} = .306$ , and significantly more positive than Black control faces ( $M = .49, SD = .30$ ),  $t(62) = 3.19, p = .002$ , Hedges'  $g_{av} = .608$ . When Frank was Black, he was also directionally more positive ( $M = .60, SD = .26$ ) than White control faces ( $M = .54, SD = .28$ ),  $t(72) = 1.58, p = .119$ , Hedges'  $g_{av} = .236$ , and Black control faces ( $M = .44, SD = .26$ ),  $t(72) = 5.22, p < .001$ , Hedges'  $g_{av} = .595$ .

Turning to a comparison of the Black and White control faces at Time 2, I found that in the control information condition, when Frank was White or Black, the Black control faces were only marginally less positive than the White control faces (both  $ps < .1$ ). In the fire rescue condition, when Frank was White, the Black control faces were significantly less positive ( $M = .49, SD = .30$ ) than the White control faces ( $M = .58, SD = .29$ ),  $t(62) = 2.58, p = .012$ , Hedges'  $g_{av} = .305$ . Likewise, when Frank was Black, the Black control faces ( $M = .44, SD = .26$ ) were significantly less positive than the White control faces ( $M = .54, SD = .28$ ),  $t(72) = 2.66, p = .01$ , Hedges'  $g_{av} = .334$ .

***Relationship to self-reported guilt.*** An examination of the exploratory measure of the self-reported extent to which participants felt guilty after reading the final information about Francis West revealed that, unsurprisingly, guilt was higher in the fire rescue condition ( $M = 3.30, SD = 1.31$ ) than in the control condition ( $M = 1.45, SD = 1.10$ ),  $F(1,278) = 164.21, p < .001$ ,  $\eta_p^2 = .371$ , with no main effect of Frank's race,  $F(1,278) < .001, p = .986, \eta_p^2 < .001$ , or interaction,  $F(1,278) = .036, p = .850, \eta_p^2 < .001$ . The guilt measure was also correlated with the size of the revision effect (A difference score of relative implicit preference for Frank over control faces at Time 2, minus the relative implicit preference at Time 1) within the fire rescue condition,  $r(134) = .321, p < .001$ , but not within the control condition,  $r(144) = -.126, p = .129$ . A mediation analysis using PROCESS (Hayes, 2013) tested the fit of guilt as a mediator of the

effect of information condition on implicit revision using 10,000 bias-corrected bootstrap samples. The 95% confidence interval for the indirect effect excluded zero,  $\beta = .1266$ , 95% CI: [.0361, .2302], Sobel  $Z = 2.98$ ,  $p = .0029$ , supporting partial mediation given that the direct effect of condition on implicit revision remained significant when controlling for guilt,  $\beta = .290$ ,  $t(279) = 4.29$ ,  $p < .001$ . It is important to note, however, that the guilt measure followed the implicit measures in the procedure, and there was also evidence to support the reverse mediation pattern (implicit revision as a mediator of the condition effect on reported guilt),  $\beta = .0657$ , 95% CI: [.0200, .1149], Sobel  $Z = 2.83$ ,  $p = .0046$ , which also left a residual direct effect,  $\beta = .544$ ,  $t(279) = 10.60$ ,  $p < .001$ .

**Explicit evaluations.** The index scores of explicit liking of Frank West were assessed in a 2 (Time: Time 1, Time 2)  $\times$  2 (Information Condition: Fire Rescue, Control)  $\times$  2 (Frank Race: White, Black) mixed-ANOVA. There was an interaction between time and information condition,  $F(1, 278) = 1020.19$ ,  $p < .001$ ,  $\eta_p^2 = .786$ , which was not moderated by Frank's race,  $F(1, 278) = .164$ ,  $p = .686$ ,  $\eta_p^2 = .001$ ; there were no significant effects of Frank's race, all  $F$ s  $< 1$ , all  $p$ s  $> .3$ .

## Discussion

The results of Study 11 provide a direct extension of the work presented in Chapter II into a societally important domain in which implicit evaluations are often characterized as difficult to reliably change: intergroup bias (Cao & Banaji, 2016; Lai et al., 2016; McConnell et al., 2008). Whereas Study 10 examined the potential for reinterpretation to produce updating in implicit impressions of a real person and found evidence for a (relatively weak) shift, Study 11 made use of the same materials employed in Chapter II to facilitate a comparison in revision effects between individuals from stigmatized vs. non-stigmatized groups. The results showed

clear support for revision in both cases: When the character was White or Black, his initial implicit negativity relative to control faces had reversed at Time 2 in the fire rescue condition. The study, however, also found support for implicit race bias, in that the revision effect was stronger when Frank was White than when he was Black; furthermore, simple effects tests showed that although Frank became more implicitly positive than Black control faces, he was only directionally more positive than White control faces. A clear race-based implicit bias persisted on the control faces, with no apparent generalization of the positive information about Frank West to the Black control faces when Frank was Black. This is consistent with a large body of findings showing how counter-stereotypical information about individual group members often fails to generalize to attitudes on the larger group, due to individuation and/or subtyping (Johnston & Hewstone, 1992; Kunda & Oleson, 1995, 1997; Queller & Smith, 2002; Richards & Hewstone, 2001; Weber & Crocker, 1983). The results do clearly show, however, that new information that provides a reinterpretation of earlier learning can result in reversals of implicit evaluations even when the person is from a group that is stereotyped as having a negative trait (hostility) consistent with an initial interpretation of individuating information. The identity dimension in question – race – is also one that is visually apparent in the prime images themselves, thus representing the kind of cue that has been hypothesized as most resistant to interference from propositional knowledge at the implicit level (McConnell et al., 2008).

One important difference between this study and Study 10, on big-game hunting, is that in the Francis West paradigm, the case for change is much more universally compelling than it is in the argument on big-game hunting: There can be little doubt that the revealed details about Francis (or Frank) West warrant a wholesale reversal in the evaluative meaning drawn from his actions. In the case of the “Rhino Hunter,” however, there is more room for doubt, counter-

arguing, motivated dismissal of the arguments, and distance between the abstract arguments and inferences about Knowlton himself. For instance, he might have other motives for hunting besides aiding conservation – like subjective enjoyment of the act – that participants might have found less positive, and the interpretation of which were not as impacted by the arguments made in the podcast. For that reason, it is not surprising that although there is evidence for the role of reinterpretation in producing implicit revision in both Study 10 and Study 11, the new information more effectively produces change in Study 11, which draws on the unambiguous Francis West story.

Collectively, Studies 10 and 11 suggest that the convincingness of new information likely matters greatly in determining the degree to which implicit impressions will be updated. However, a remaining question regarding the application of this work on implicit updating is whether new information that is ostensibly just as clear-cut and unambiguous as the fire rescue details, but less rare and extreme, can also effectively produce revision. Although revision after strongly heroic actions or thorough argumentation may be possible, it is much more common in daily life for corrections in our construal of others to be quick and more minor and routine. How effectively can new information that updates knowledge of others in a more mundane way revise implicit impressions? Study 12 turns to this question.

### **Study 12: Counter-Stereotypical Impressions – II**

In recent work on the formation of counter-stereotypical person impressions, Cao and Banaji (2016) assessed implicit impressions of a novel man named Jonathan and a novel woman named Elizabeth as being relatively more associated with the concept “doctors” or “nurses”. They found that prior to individuating information, participants had a stronger implicit impression of Jonathan as a doctor and Elizabeth as a nurse than of Jonathan as a nurse and

Elizabeth as a doctor (the measure used – the IAT – allows only for statements about *relative* impression strength). However, after learning that Elizabeth was in fact the doctor, and Jonathan the nurse, the implicit impressions had shifted but *not* reversed; they remained significantly in the stereotype-consistent direction, even though participants explicitly reported that they now believed Elizabeth to be the doctor and Jonathan the nurse, consistent with what they had learned. The new information thus seemed to impact explicit judgments much more strongly than implicit impressions, despite the observation that, like the reinterpretation procedures used throughout the work presented in this chapter and the preceding one, these clarifications involve a seemingly straightforward and compelling reason to shift one’s construal of Elizabeth and Jonathan. In fact, this seems much like the “negate + affirm” strategy in Studies 7-8, in which an initial interpretation is invalidated while being replaced by another; here, Elizabeth being revealed as the doctor and Jonathan as the nurse clearly negates the initial stereotype-consistent assumption and replaces it with counterstereotypical information that is asserted to be true. Why might revision be weaker in this case, when the prior studies of Chapters II-III have found robust evidence for implicit revision?

One possibility why revision might be weaker in the work of Cao and Banaji (2016), as previewed in the discussion for Study 11, is that new information may need to be more extreme in order to prompt large-scale revision. After all, extremity is a generally a strong cue to diagnosticity (Skowronski & Carlston, 1989), and information that is diagnostic by virtue of its extremity has been demonstrated to strongly impact implicit evaluations (Cone & Ferguson, 2015). Of course, extremity may not be the only cue that information is diagnostic of the character of a person, and a clear statement that Elizabeth is in fact a doctor would seem on its face to reveal something relevant and diagnostic about her, as Cao and Banaji (2016) argued.

This might entail that even if new information is diagnostic and subjectively believed by participants, this alone may be insufficient to produce full revision at the implicit level. It may be, for example, that integration of the new information into implicit responses requires a stronger expectancy violation, which more extreme (or surprising) information delivers. If so, this would limit the range of circumstances under which relevant new information about a person can lead to rapid updating of implicit impressions.

A possible middle ground between these alternatives (that diagnostic information is always effective vs. only effective when it is extreme) is that there may be moderators that “boost” the effectiveness of more mundane diagnostic information in enacting implicit revision. For example, a more self-relevant or immersive scenario during learning may allow for implicit updating with less extreme forms of relevant new information about a novel person.

Perspectives on the various forms of processing and systems involved in implicit social cognition have noted that the formation and change of responses in the context of instrumental learning is often much faster than other forms of learning (Amodio & Ratner, 2011), and implicit processes can be quickly calibrated by personal goals (Ferguson & Wojnowicz, 2011), fitting the idea that there is greater impetus to accurately process and perceive stimuli that are directly relevant to the actions and outcomes of a perceiver. If Elizabeth and Jonathan are seen as abstract, vague individuals who the participant never expects to meet, information about them may be processed deeply enough to accurately recall information about them when prompted, but not deeply enough to overcome deeply ingrained stereotype-based responses. Though participants also do not expect to meet Francis West, his story is more detailed, immersive, and engaging than the minimal information about Jonathan and Elizabeth; furthermore, participants are told that the story is based on real events. If Jonathan and Elizabeth can be portrayed as more concrete,

relevant people, on the other hand, then perhaps counter-stereotypical information about their professions will prove more impactful – conditions which the comparably elaborate materials in the earlier studies in this package may have produced. The current study will begin to test this possibility by manipulating whether the information presented about the novel individuals is relatively minimal (following the materials used by Cao & Banaji, 2016) or more immersive and self-relevant.

Another possible reason for the comparably weak revision found by Cao and Banaji (2016) pertains to the implicit measure that was used in those studies, the Implicit Association Test (IAT; Greenwald et al., 1998). Because the IAT requires participants to sort stimuli appearing on the screen into categories, evidence suggests that it is often more effective at measuring implicit impressions of the overall categories rather than particular stimuli themselves, because the stimuli are construed in terms of the salient categories (De Houwer, 2001). Because Cao and Banaji (2016) used names as the category labels (“Elizabeth” and “Jonathan”) and nicknames as stimuli (e.g., eliza, ell, jon, johnny) – which had not actually been used in the information conveying the professions of the *particular* Jonathan and Elizabeth being learned about in the study – it is possible that this test captured more general information about the participants’ implicit impressions of the broader categories of Jonathans and Elizabeths rather than solely implicit impressions of *this* Jonathan and *this* Elizabeth. Though Cao and Banaji (2016) attempted to address this concern in part in a follow up study (Study 3) using targets with the novel names Lapper and Affina, with novel nicknames for these as stimuli, the clearly gendered nature of these names means that it continues to be possible that an IAT using these names as categories would tap into the broader gender associations that these names connote, instead of (or in addition to) the individuating information previously presented about *this*



Lapper and *this* Affina. To address this possibility, the current study will modify the IAT to present face images of Elizabeth and Jonathan as the category labels, as well as compare responses on this measure to a modification of the AMP that is not designed to prompt participants to construe the prime stimuli as members of a shared category.

## **Method**

**Participants.** To obtain responses from at least 100 participants in each of 8 between-subjects conditions of the study, I aimed to recruit 800 participants on Mechanical Turk, and received complete response from 803. Of these, the following number were excluded a priori: 7 due to loss of data due to server timeout, 4 due to familiarity with Mandarin and/or Cantonese in the PMP condition, 1 for using a single key on every trial of the PMP, 24 for responding more quickly than 300ms on at least 10% of trials in the IAT condition, and 18 for failing an attention check. This left 749 participants for analysis (52% women, Age  $M = 37.6$  years,  $SD = 12.1$ ).

**Initial learning.** Participants first saw a screen of information similar to that used by Cao and Banaji (2016), in which they were introduced to two novel people: Jonathan and Elizabeth. Unlike in Cao and Banaji's (2016) work, the present study, these introductions were accompanied by face photographs (for use on the subsequent implicit measures). For each participant, an image of Elizabeth and an image of Jonathan were randomly drawn from a set of 5 female and 5 male photographs, respectively. These sets were selected from a larger pool of stimuli in the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015) by a custom algorithm that calculated the mean squared difference for 10,000 samples of randomly-drawn sets of 5 female and 5 male faces across 12 pre-rated dimensions (afraid, angry, attractive, babyfaced, dominant, disgusted, happy, sad, surprised, threatening, trustworthy, unusual) and then selected

the set of 10 with the lowest difference score between sexes, so as to minimize differences between the male and female face sets as much as possible.

Following Cao and Banaji (2016), participants were informed that Jonathan and Elizabeth were gainfully employed in a large city, with one of the two individuals having a job as a doctor, and the other having a job as a nurse. It was made clear to participants that they did not yet know which person held which position. The names and images of both participants were presented on the same page, with left/right position counterbalanced.

**Additional information about Jonathan and Elizabeth.** Following the initial information screen, participants learned individuating information about the professions of Jonathan and Elizabeth. They were assigned to either the minimal or immersive materials conditions, and the stereotypical or counter-stereotypical information condition.

***Minimal condition.*** The minimal information condition followed the procedure of Cao and Banaji (2016). Participants viewed a single screen of information about the profession of Jonathan, and a single screen of information about the profession of Elizabeth, in counterbalanced order. Each screen included the photo of the target person. In the stereotypical condition, Jonathan was revealed to be the doctor and Elizabeth the nurse; in the counter-stereotypical condition, Elizabeth was revealed to be the doctor and Jonathan the nurse. The descriptions of the two professions are provided below (text that varies is presented in brackets):

[Jonathan/Elizabeth] is a doctor at a city hospital where [he/she] specializes in emergency medicine. As a physician working in an emergency room, [Jonathan/Elizabeth] cares for patients who arrive at the hospital requiring immediate medical attention. Since patients arrive with a variety of ailments, [Jonathan/Elizabeth] is trained in resuscitation, cardiac life support, airway management, and some surgical procedures. After stabilizing patients, [Jonathan/Elizabeth] decides whether to release them or admit them to the hospital for further treatment.

[Jonathan/Elizabeth] is nurse at a retirement home where [he/she] provides care to the elderly. As a nurse working at a retirement home, [Jonathan/Elizabeth] cares for the elderly who get sick or hurt unexpectedly and for the elderly with chronic ailments. Given these different needs, [Jonathan/Elizabeth] is trained to administer medications, bandage wounds, and monitor blood pressure, heart rates and respiration. After treating the elderly, [Jonathan/Elizabeth] decides whether to release them or send them to a hospital for additional care.

***Immersive condition.*** In the immersive condition, participants were informed that they would be reading a description of events, and that they should imagine that the events were actually occurring, and happening to them. They were directly asked to attempt to “immersive” themselves in the situations described, and told that they will be asked questions about the information later.

On the following pages, the participants read a story in the second person (using “you” as pronoun) in which they are driving home after work, and suddenly experience searing pain in their side. It is bad enough to force them to pull over and call an ambulance in a panic (the full materials are available in the Appendix). After reading that they arrive at the hospital, the participants were presented with a screen showing the images of Jonathan and Elizabeth, with the following text to inform them of which person filled which role (text that varies presented in brackets): *Once you get to the hospital, you meet your nurse, [Jonathan/Elizabeth], and lead doctor, [Elizabeth/Jonathan].* The two individuals were presented side-by-side, in counterbalanced position.

The story continued on a final page that also featured the names and images of Jonathan and Elizabeth, with the following text:

They read the chart from the ambulance medics, and rush you into an exam room to take further tests. [Jonathan/Elizabeth], the nurse, is taking your vitals again, and [Elizabeth/Jonathan], the doctor, is

deciding which course to follow. You will need to ask them questions and so you should take a minute to remember their face and name carefully and their role in diagnosing your condition.

**Explicit measure.** After reading either the minimal or immersive materials about Elizabeth and Jonathan, participants were asked the single item measure from Cao and Banaji's work (2016) to assess the explicit beliefs of the participant about the professions of Jonathan and Elizabeth, selecting a response on a 7-point Likert-type scale from -3 (*Jonathan is definitely the doctor*) to +3 (*Elizabeth is definitely the doctor*), with the midpoint (0) labeled *Both individuals are equally likely to be the doctor*.

**Implicit measure.** Participants next completed either an IAT or modified AMP (described below), both designed to measure relative implicit impressions of Jonathan vs. Elizabeth as more linked with doctors or nurses. Though Cao and Banaji (2016) administered the implicit measure twice, once before the individuating information (to establish a stereotype-consistent baseline) and once after the individuating information, in this study there was only a single administration, after the final information was presented. This was done to keep the study shorter for reasons of cost, and because the focus was on whether counter-stereotypical implicit impressions could form with these materials, with the work of Cao and Banaji (2016) already establishing the baseline stereotypicality of implicit impressions prior to individuating details.

***Implicit Association Test.*** The IAT required participants to sort stimuli related to Jonathan, Elizabeth, doctors, and nurses, appearing sequentially in the center of the screen, into four categories using two keys. In the critical blocks of the experiment, there were two categories mapped to each key, the configuration of which varied between blocks: In the stereotype-compatible block, Jonathan + doctor were mapped together on one key and Elizabeth + nurse were mapped together on the other key, and on the stereotype-incompatible block, Jonathan + nurse were mapped together on one key and Elizabeth + doctor were mapped

together on the other key. The images of Jonathan and Elizabeth were displayed on-screen as the category labels, with the trial stimuli to be categorized including *Jonathan, jon, John, johnny*, and *Elizabeth, ell, Ella, eliza*, respectively. The doctor and nurse categories were represented by the labels “Doctor Words” and “Nurse Words”, with trial stimuli to be categorized including *Doctor, Medical Doctor, M.D., Physician*, and *Nurse, Registered Nurse, R.N., Nursing*, respectively. The IAT consisted of 7 blocks (including practice and test blocks) with the standard 20-20-20-40-40-20-40 trial structure (see Greenwald et al., 2003). Participants were instructed to go as quickly as they could without making many mistakes. Upon an error, participants were shown a red “X” and were required to press the correct key before advancing to the next trial.

***Profession Misattribution Procedure.*** The AMP has attractive psychometric properties to rival the IAT, including high reliability, lack of reliance on response times (which are measured with error), lack of the block-order effects present on the IAT (as whichever blocks are presented second, be they either the compatible or incompatible blocks, generally suffer from interference effects from the mappings learned in the first half of the task). In addition, and particularly relevant for the present study, the AMP does not impose any category-level construal of the stimuli; in fact, the instructions on the AMP generally ask participants to do their best to disregard the primes entirely. This may make an AMP preferable for measuring implicit impressions of individuals (vs. the broader categories of “Jonathan” and “Elizabeth”) compared to the IAT.

The AMP in its standard form, of course, is not structured to measure implicit impressions of a person as a doctor or nurse, because the decision that participants make about each pictograph pertains to its valence – i.e., whether the pictograph is more or less pleasant than average. It is an *affective* misattribution task, such that evaluations spontaneously evoked by the

primes are misattributed to the pictographs that are the participant's focus of attention. Work suggests, however, that a variety of semantic impressions can be misattributed, including impressions of animacy (Deutsch & Gawronski, 2009), self-relevance (Sava et al., 2012), sexual interest (Imhoff et al., 2011), and racial stereotypes (Krieglmeyer & Sherman, 2012). As another example, in other work, I have found evidence that implicit impressions of the *creativity* of a person can be measured on an AMP-like misattribution task in which images of people are primed directly before abstract painting stimuli that participants must judge as more or less creative than average (the stimuli used in Study 10). These implicit judgments were distinct from implicit evaluations measured on a standard AMP and uniquely predicted judgments and behavioral intentions when controlling for explicit judgments of the target's creativity (Mann, Katz, Ferguson, & Goncalo, in preparation). Such work raises the possibility that a focal task aimed at determining the "doctor-ness" or "nurse-ness" of the pictographs could potentially tap misattributions of impressions of Jonathan and Elizabeth as doctors or nurses, thereby providing an implicit index of such impressions.

At the beginning of this "Profession" Misattribution Task (PMP), participants were told that they would be judging a number of pictographs, with some of them having a meaning more related to "doctors" and others having a meaning more related to "nurses." We gave them a cover story that the meaning of such pictographs can often be gleaned from their appearance, and that people are often remarkably good at detecting this. They were thus asked to press one key if the meaning of a pictograph seemed to them to be more related to doctors, and a different key if they thought the meaning was more related to nurses, with the usual AMP instruction to avoid being biased at all by the prime images that came before each. As with the regular AMP, each

trial consisted of the presentation of a prime for 75ms, a blank page for 125ms, a pictograph for 100ms, and finally a black-and-white pattern mask until the participant gave a response.

Because the PMP allows for more than two prime categories, in addition to Jonathan and Elizabeth the task included control faces as well, to allow an assessment of how implicit impressions of Jonathan and Elizabeth in each condition compared to control individuals of the same or opposite sex. There were thus 20 trials each with the following stimuli as primes: Elizabeth, Jonathan, control women (divided evenly among the four unused images of women), and control men (divided evenly among the four unused images of men), for a total of 80 trials.

**Survey questions.** Participants next completed a number of questionnaire items asking about various elements of the study, in the order of the following sections. The goal of most of these questions was to determine a) the degree to which implicit impressions formed in this study might relate to important judgments about the targets of those impressions, and b) whether responses to the questions might be impacted by the information and/or materials manipulations in a way that could shed light on the inferences that participants are making in each condition, ultimately informing why and how implicit impression formation might be stronger or weaker. The analyses planned for these questions were thus largely exploratory in nature.

**Believability.** Four exploratory questions assessed the degree to which participants found aspects of the procedure believable. These included, “As you were going through the experiment, how much did you think of the people you learned about as real-life people?” on a scale from 1 (*Not at all*) to 9 (*Completely*), “Do you think the people you learned about—Jonathan and Elizabeth—are actually real people or just made-up fictional examples?” on a scale from 1 (*Definitely fictional people*) to 9 (*Definitely real people*), “Do you think the descriptions you read about Jonathan and Elizabeth are realistic?” on a scale from 1 (*Not at all realistic*) to 9

(*Completely realistic*), and “Do you think the descriptions of these people are just made-up fictional descriptions that the experimenters are using for their own purposes?” on a scale from 1 (*Not at all*) to 9 (*Definitely*).

**Attention check.** To identify participants who were not paying adequate attention to the questions given the large set of items in the questionnaire, the four believability questions above were divided between the 2<sup>nd</sup> and 3<sup>rd</sup> by an attention check, consisting of a paragraph ostensibly asking participants to insert “a fact about Jonathan” and “a fact about Elizabeth” in two boxes below (bearing those phrases as labels). In fact, however, the penultimate sentence instructed them to disregard the rest of the question and simply type the letter “z” in each box.

**Identity importance.** Two questions measured how much participants believed the professions of Elizabeth and Jonathan to be important to their self-identities. Each was asked twice, once for Elizabeth and once for Jonathan: “To what extent do you believe that [Elizabeth’s/Jonathan’s] profession is an important part of who [she/he] is as a person?” on a scale from 1 (*Not at all*) to 9 (*Extremely*), and “To what extent do you believe that [Elizabeth/Jonathan] views [her/his] profession as an important part of [her/his] identity?” on a scale from 1 (*Not at all*) to 9 (*Extremely*).

**Competence.** In three questions, participants indicated how competent they viewed Jonathan and Elizabeth to be. As with the identity importance questions, each was asked twice, once for Jonathan and once for Elizabeth: “If you had to guess, without knowing anything else, how competent do you think [Jonathan/Elizabeth] is at [his/her] job?” on a scale from 1 (*Not at all competent*) to 9 (*Extremely competent*), “If [Jonathan/Elizabeth] were involved in caring for a member of your family, to what degree do you feel that your family member would be ‘in good hands,’ based on what you know now?” on a scale from 1 (*Not at all*) to 9 (*Extremely*), and



finally, “If you had to guess, what is the likelihood that [Jonathan/Elizabeth] was valedictorian of [his/her] graduating class for [his/her] program?” on a scale from 1 (*Extremely unlikely*) to 9 (*Extremely likely*).

**General views on gender stereotypes.** Participants indicated whether they believed that a randomly selected man or a randomly selected woman from the medical industry would be more likely to be a doctor vs. a nurse, to the best of their knowledge (and without taking into account what they know about Jonathan and Elizabeth), using a scale from 1 (*The woman is more likely to be the doctor*) to 9 (*The man is more likely to be the doctor*). They were then asked their level of agreement with the statement, “In general, people tend to assume on average that doctors are male and that nurses are female” on a scale from 1 (*strongly disagree*) to 9 (*strongly agree*). Finally, they were asked how a random person on the street would respond to a question on whether someone named Elizabeth or Jonathan is more likely to be a doctor, and could select one of three answers for this hypothetical person: *Elizabeth is more likely to be the doctor*, *Jonathan is more likely to be the doctor*, or *They are equally likely to be the doctor*.

**Other questions.** Lastly, participants were asked the degree to which they tried to immerse themselves in and vividly imagine the details about Jonathan and Elizabeth on a scale from 1 (*Not at all*) to 9 (*Completely*), to identify both Elizabeth and Jonathan in photo lineups of the 5 images of women and 5 images of men, to indicate if they knew Mandarin and/or Cantonese (in the PMP condition), and provide demographic information.

## Results

**Implicit beliefs.** Implicit beliefs about the degree to which Elizabeth and Jonathan were linked with doctors and nurses were assessed using either the IAT or PMP, depending on condition assignment.

**Profession Misattribution Procedure.** For participants in the PMP condition, I computed the proportion of pictographs judged to be more related to doctors (vs. nurses) separately for each prime type, and analyzed these proportions within a 4 (Prime Type: Jonathan, Elizabeth, control men, control women) x 2 (Information Condition: stereotypical, counter-stereotypical) x 2 (Materials: minimal, immersive) mixed ANOVA, with prime type manipulated within-participants and the latter two factors manipulated between-participants.

Of main importance, the interaction between person prime and information condition was significant,  $F(1.53, 574.69) = 86.30, p < .001, \eta_p^2 = .187$ . Figure 16 shows implicit evaluations of each of the four prime types within each condition. This interaction was *not* qualified by materials condition (minimal vs. immersive) in a three-way interaction,  $F(1.528, 574.69) = .56, p = .526, \eta_p^2 = .001$ , so interpretation will focus on the two-way effect.

In the stereotypical information condition, in which Jonathan was revealed to be the doctor and Elizabeth the nurse, Jonathan was more implicitly linked with doctors (vs. nurses;  $M = .69, SD = .22$ ) than Elizabeth ( $M = .39, SD = .25$ ),  $t(189) = 9.47, p < .001$ , Hedges'  $g_{av} = 1.255$ . Jonathan was also judged as more implicitly doctor-like than control men ( $M = .65, SD = .21$ ),  $t(189) = 2.57, p = .011$ , Hedges'  $g_{av} = .153$ , and control women ( $M = .39, SD = .23$ ),  $t(189) = 10.23, p < .001$ , Hedges'  $g_{av} = 1.314$ . Implicit judgments of Elizabeth did not differ from implicit judgments of control women,  $t(189) = .061, p = .952$ , Hedges'  $g_{av} = .003$ , but Elizabeth was implicitly judged as less doctor-like than control men,  $t(189) = 9.15, p < .001$ , Hedges'  $g_{av} = 1.124$ . Finally, control women were implicitly judged as less doctor-like than control men,  $t(189) = 9.50, p < .001$ , Hedges'  $g_{av} = 1.177$ .

In the counter-stereotypical condition, in which Elizabeth was the doctor and Jonathan was the nurse, Elizabeth was more implicitly linked with doctors (vs. nurses;  $M = .63, SD = .24$ )

than Jonathan ( $M = .45$ ,  $SD = .25$ ),  $t(189) = 5.658$ ,  $p < .001$ , Hedges'  $g_{av} = .730$ . Elizabeth was also implicitly judged as more doctor-like than control women ( $M = .56$ ,  $SD = .22$ ),  $t(189) = 4.16$ ,  $p < .001$ , Hedges'  $g_{av} = .296$ , and control men ( $M = .49$ ,  $SD = .23$ ),  $t(189) = 4.55$ ,  $p < .001$ , Hedges'  $g_{av} = .580$ . Additionally, Jonathan was implicitly judged as less doctor-like (and thus more nurse-like) than control men,  $t(189) = 3.09$ ,  $p = .002$ , Hedges'  $g_{av} = .178$ , and control women,  $t(189) = 3.89$ ,  $p < .001$ , Hedges'  $g_{av} = .469$ . Finally, control women were implicitly judged as more doctor-like than control men,  $t(189) = 2.39$ ,  $p = .018$ , Hedges'  $g_{av} = .302$ .

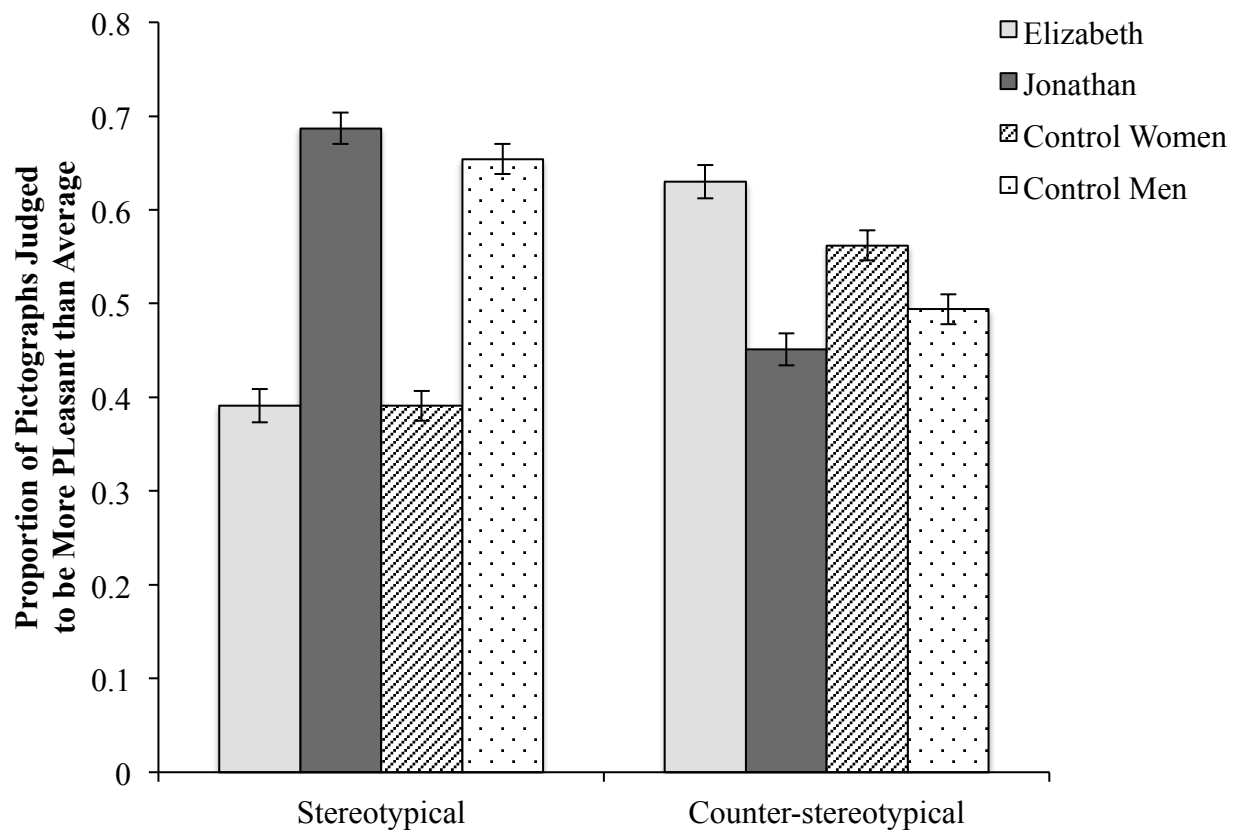


Figure 16. The proportion of pictographs judged to be more pleasant than average in Study 12 (PMP condition), by information condition and prime. Error bars are standard errors.

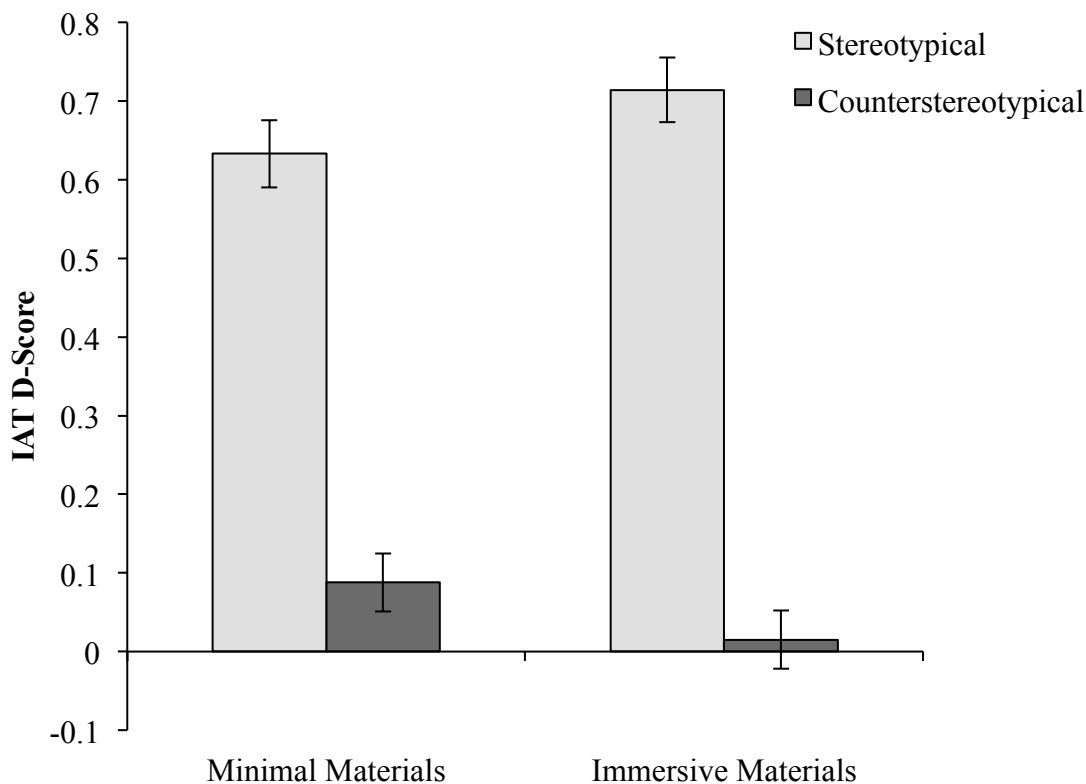
***Implicit Association Test.*** IAT scores were computed using the D1-scoring algorithm (Greenwald et al., 2003), which takes advantage of the built-in error penalties available in this

version of the IAT that requires participants to correct their own errors by pressing the correct response key. These *D* scores are relative measures interpretable as standardized mean differences, bounded between -2 and +2. As noted in the Method section, participants were excluded if they responded more quickly than 300ms on 10% or more trials. Positive scores indicate stronger implicit judgments of Jonathan as a doctor and Elizabeth as a nurse, and negative scores indicate stronger implicit judgments of Elizabeth as a doctor and Jonathan as a nurse. *D*-scores were analyzed in a 2 (Information Condition: stereotypical, counter-stereotypical) x 2 (Materials: minimal, immersive) between-participants factorial ANOVA.

This analysis revealed a main effect of information condition,  $F(1, 365) = 249.57, p < .001, \eta_p^2 = .406$ , qualified by a (marginal) interaction with materials condition,  $F(1, 365) = 3.84, p = .051, \eta_p^2 = .01$ . The nature of this interaction was such that the effect of individuating information about the professions of Jonathan and Elizabeth on IAT scores was stronger in the immersive materials condition,  $F(1, 365) = 161.10, p < .001, \eta_p^2 = .306$ , than in the minimal materials condition,  $F(1, 365) = 93.74, p < .001, \eta_p^2 = .204$ . Figure 17 shows the mean *D*-scores in each information and materials condition.

In the minimal materials condition, when the information about Jonathan and Elizabeth was stereotype-consistent, *D*-scores were consistent with Jonathan being implicitly judged as more doctor-like (vs. nurse-like) than Elizabeth ( $M = .63, SD = .33$ ), which was significantly above zero,  $t(76) = 16.60, p < .001$ . When the information was counter-stereotypical, *D*-scores still suggested that Jonathan was implicitly judged as more doctor-like (vs. nurse-like) than Elizabeth ( $M = .09, SD = .41$ ), albeit less so, and this too was significantly above zero,  $t(104) = 2.21, p = .029$ .

In the immersive materials condition, when the information about Jonathan and Elizabeth was stereotype-consistent, *D*-scores were consistent with Jonathan being implicitly judged as more doctor-like (vs. nurse-like) than Elizabeth ( $M = .71$ ,  $SD = .33$ ), which was significantly above zero,  $t(84) = 19.69$ ,  $p < .001$ . When the information was counter-stereotypical, *D*-scores still suggested that neither Jonathan nor Elizabeth was implicitly judged as more doctor-like (vs. nurse-like;  $M = .02$ ,  $SD = .40$ ), as the mean *D*-score in this condition did not significantly differ from zero,  $t(101) = .38$ ,  $p = .702$ .



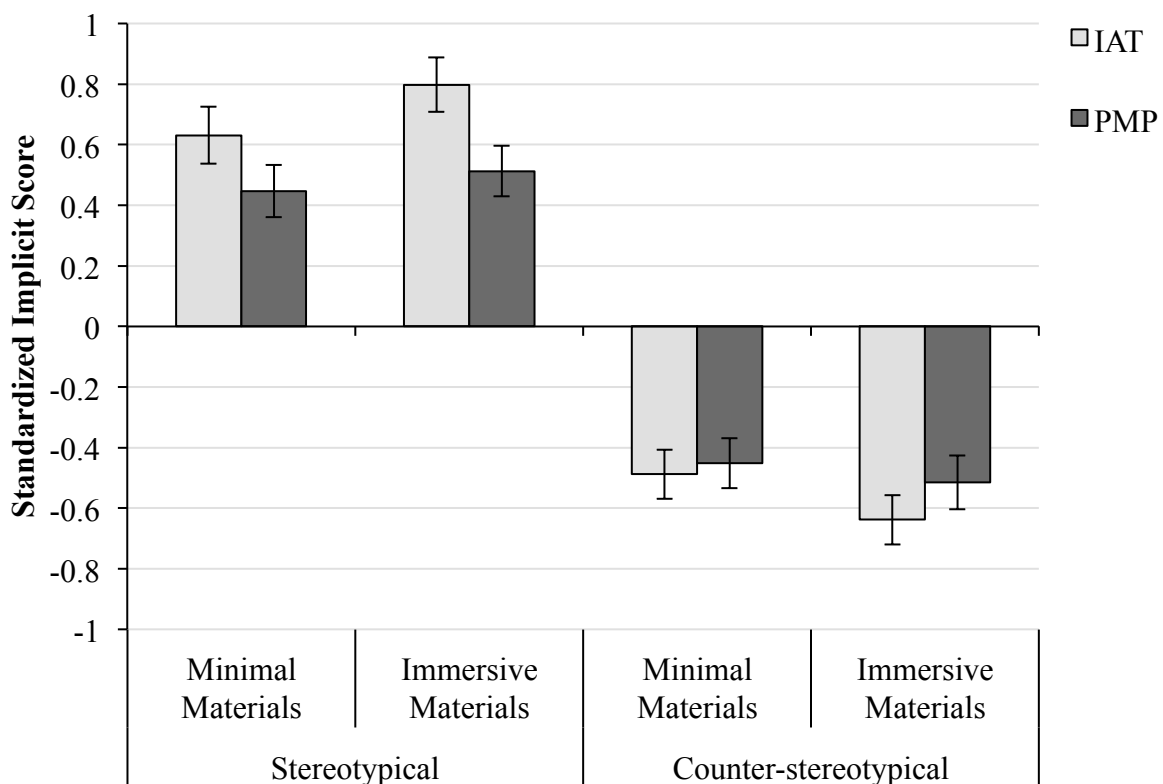
*Figure 17.* Mean *D*-score index of relative implicit judgment of Jonathan as doctor and Elizabeth as nurse in Study 12 (IAT condition), by information condition and materials condition. Error bars are standard errors.

***Combined analysis of PMP and IAT.*** After examining the effects of information condition and material immersiveness separately for the PMP and IAT, a natural next question is to ask how these effects compare between the measures. Is the size of the information effect similar on the two measures? Is the suggestive impact of materials on the condition effect as measured by the IAT a significant difference from the PMP? A comparison of the two measures may be useful given that the difference in pattern between them, especially the reversal of implicit judgments on the PMP vs. the mere shift to neutral of implicit stereotyping on the IAT, gives an impression of a fundamental dissimilarity between these measures. However, they may react similarly to the manipulations (information and materials) and produce the aforementioned divergence in means due to an influence exogenous to those manipulations, like relative sensitivity to social group memberships of the individuals presented during the task. Specifically, because the IAT requires participants to think about stimuli in terms of categories, it might incorporate category stereotypes to a greater degree than the PMP, which does not impose this task framing. This would result in overall higher levels of implicit gender stereotyping on the IAT, as was observed here. However, the IAT might *also* draw upon individuating information to the same extent as the PMP. A direct comparison between the measures can allow for a test of this idea.

In order to compare these measures that have very different scales, I carried out two transformations. First, because the IAT is a relative measure of implicit judgments of Jonathan vs. Elizabeth, I created a difference score on the PMP by subtracting the proportion of Jonathan trials on which the pictograph was judged as more doctor-related minus the proportion of Elizabeth trials on which the pictograph was judged as more doctor-related. Following this, both the IAT scores and PMP difference scores were standardized separately, and then combined into

a single measure. These scores were then assessed in a 2 (Information Condition: stereotypical, counter-stereotypical) x 2 (Materials: minimal, immersive) x 2 (Measure: IAT, PMP) between-participants factorial ANOVA. Figure 18 displays the mean standard scores within each condition.

Results showed a significant and expected main effect of information condition,  $F(1, 741) = 339.75, p < .001, \eta_p^2 = .314$ . This was qualified by a significant interaction between information and implicit measure,  $F(1, 741) = 6.68, p = .010, \eta_p^2 = .009$ , and a marginal interaction between information and materials,  $F(1, 741) = 3.35, p = .068, \eta_p^2 = .004$ . There was no three-way interaction between information, materials, and implicit measure,  $F(1, 741) = .604, p = .437, \eta_p^2 = .001$ .



*Figure 18.* Standardized implicit judgment of Jonathan as more doctor-like (vs. nurse-like) than Elizabeth in Study 12, by information condition, materials, and implicit measure. Error bars are standard errors.

Examining the interaction between information condition and measure revealed that the measures significantly differed in the stereotypical information condition, but not in the counter-stereotypical condition. In the stereotypical condition, the IAT showed stronger implicit judgments of Jonathan as relatively more doctor-like (vs. nurse-like) than Elizabeth ( $M = .72$ ,  $SD = .69$ ) compared to the PMP ( $M = .48$ ,  $SD = .87$ ), unequal variances  $t(348.09) = 2.80$ ,  $p = .005$ , Hedges'  $g_s = .299$ . In the counter-stereotypical condition, however, the IAT showed no stronger relative implicit judgments of Jonathan as the doctor ( $M = -.56$ ,  $SD = .83$ ) compared to the PMP ( $M = -.48$ ,  $SD = .88$ ),  $t(395) = -.95$ ,  $p = .345$ , Hedges'  $g_s = .095$ .

Turning now to the marginal interaction between information and materials condition ( $p = .068$ ), follow-up tests indicated that, as with the IAT analysis, the effect of information condition was larger with the immersive materials,  $t(370) = 14.13$ ,  $p < .001$ , Hedges'  $g_s = 1.46$ , than with the minimal materials,  $t(375) = 11.74$ ,  $p < .001$ , Hedges'  $g_s = 1.21$ . The nonsignificant moderation of this effect by implicit measure ( $p = .437$ ) suggests that this trend does not differ between the IAT and the PMP.

**Explicit judgments.** The measure asking participants to indicate their view about whether they believed Jonathan or Elizabeth to be the doctor was analyzed in a 2 (Information Condition: stereotypical, counterstereotypical) x 2 (Materials: minimal, immersive) between-participants ANOVA, and revealed only a main effect of information,  $F(1, 745) = 3873.67$ ,  $p <$



.001,  $\eta_p^2 = .839$ . There was no main effect of materials,  $F(1, 745) = .011, p = .918, \eta_p^2 < .001$ , or interaction,  $F(1, 745) = .848, p = .357, \eta_p^2 = .001$ .

**Inference questions.** The final questions in the experiment were organized broadly into the following categories: (a) the perceived importance of the careers of Jonathan and Elizabeth to their identities, (b) the perceived competence of Jonathan and Elizabeth, (c) predictions on how other people in general would view Jonathan and Elizabeth, and average levels of gender-based stereotypes of the doctor and nurse professions among people in general, and (d) general questions about the believability and immersiveness of the study. For each group of questions, I briefly report whether they were affected by the information and/or materials manipulations, and whether they were related to implicit and explicit beliefs. Due to the exploratory nature of these items, these analyses are kept more concise than the main analyses reported above, and analyses of questions falling into the fourth category (believability and immersion) are not reported.

**Importance of career to identity.** Four questions dealt with the degree to which participants believed that Jonathan and Elizabeth viewed their jobs as important parts of their identity (two each). The two questions on Jonathan's perceived identification with his profession were highly correlated,  $r(747) = .77, p < .001$ , as were the two similar questions regarding Elizabeth,  $r(747) = .81, p < .001$ ; these pairs were thus averaged to create single measures of identification per person. As the implicit measures are relative (in the standardized form computed above to allow their comparison), in order to examine the relationship between the implicit measures and the career importance measures, a career importance difference score was produced by subtracting the perceived importance of Elizabeth's career from the perceived importance of Jonathan's (with higher scores thus indicating a perception that Jonathan's career is more important to his identity).

First, the difference score were significantly affected by information condition, but not the materials manipulation (and there were no interactions); the relative perceived identity importance of Jonathan's career (vs. Elizabeth's) was higher when Jonathan was the doctor ( $M = .15$ ,  $SD = .79$ ) than when Elizabeth was the doctor ( $M = -.41$ ,  $SD = .87$ ),  $t(747) = 9.23$ ,  $p < .001$ , Hedges'  $g_s = .675$ . In other words, the perceived importance of each person's career to their identity was greatest when he/she was the doctor rather than the nurse.

This difference score was regressed upon the combined standardized implicit measure and information condition, which revealed a significant unique effect of implicit beliefs,  $\beta = .086$ ,  $t(746) = 2.07$ ,  $p = .039$ , such that higher implicit belief that Jonathan (vs. Elizabeth) was doctor-like (vs. nurse-like) predicted a greater explicit belief that Jonathan's job is important to his identity. This was no longer significant when additionally controlling for the single item of explicit belief that Jonathan (vs. Elizabeth) was the doctor,  $\beta = .058$ ,  $t(745) = 1.39$ ,  $p = .165$ .

***Perceived competence.*** For each of Jonathan and Elizabeth, three questions asked about their competence (how competent they are, whether one of the participant's own family members would be in good hands under their care, and the likelihood that they were valedictorian of their graduating class). These three items were reliably inter-correlated for Jonathan (Cronbach's  $\alpha = .723$ ) and Elizabeth ( $\alpha = .708$ ), and were thus merged into a single index score for each target individual. Similar to the analysis for perceived identity importance, a difference score was created by subtracting the score for Elizabeth from the score for Jonathan, with higher scores indicating greater perceived competence of Jonathan (vs. Elizabeth).

Relative perceived competence of Jonathan over Elizabeth was significantly impacted by information condition,  $F(1,745) = 92.65$ ,  $p < .001$ ,  $\eta_p^2 = .111$ , as well as materials,  $F(1,745) = 5.99$ ,  $p = .015$ ,  $\eta_p^2 = .008$ , and the interaction between the two factors,  $F(1,745) = 8.35$ ,  $p = .004$ ,

$\eta_p^2 = .011$ . The relative perceived competence of Jonathan over Elizabeth was greater when he was the doctor than when he was the nurse in both the minimal materials condition,  $t(375) = 8.11, p < .001$ , Hedges  $g_s = .838$ , and the immersive materials condition,  $t(370) = 5.30, p < .001$ , Hedges  $g_s = .549$ , but this difference was greater in the former than the latter. This pattern indicates that the competence judgments of each individual were higher when he/she was the doctor rather than the nurse.

Implicit impressions of Jonathan vs. Elizabeth were marginally related to relative competence judgments when controlling for information, materials, and their interaction,  $\beta = .075, t(744) = 1.82, p = .070$ , but not when additionally controlling for explicit beliefs,  $\beta = .057, t(743) = 1.37, p = .173$ .

***General beliefs about the inferences of others in general.*** Because the three items in this category were less strongly inter-correlated and were answered using scales of different lengths, they were analyzed separately.

The question asking participants to infer whether a random man or random woman working in the medical industry would be more likely to be a doctor rather than a nurse (9-point scale) was not affected by information condition, materials condition, or their interaction, all  $F_s < 2$ , all  $p_s > .164$ . When the combined standardized implicit measure was added to the analysis, however, it had a significant, unique positive effect,  $\beta = .155, t(744) = 3.54, p < .001$ , such that greater implicit judgment of Jonathan (vs. Elizabeth) as doctor-like (vs. nurse-like) was associated with greater belief that a random man from the medical industry would be more likely to be a doctor than a random woman in that industry. This predictor remained significant,  $\beta = .163, t(743) = 3.66, p < .001$ , when controlling for explicit beliefs in the likelihood of Jonathan

being the doctor over Elizabeth, which was not independently significant,  $\beta = .091$ ,  $t(743) = .995$ ,  $p = .320$ .

The question on whether “people in general” tend to assume that doctors are male and nurses are female (9-point scale) was not affected by the manipulations (or their interaction), all  $F_s < 3$ , all  $p_s > .1$ . With the combined standardized implicit measure added to the analysis, a marginally significant positive trend obtained,  $\beta = .083$ ,  $t(744) = 1.89$ ,  $p = .059$ , which became significant,  $\beta = .094$ ,  $t(743) = 2.10$ ,  $p = .036$ , when additionally controlling for explicit judgments of the professions of Jonathan and Elizabeth. Explicit judgments were not predictive,  $\beta = .121$ ,  $t(743) = 1.31$ ,  $p = .190$ .

Last was the question on which option a random person on the street would select: a) that someone named “Elizabeth” was more likely to be a doctor than someone named “Jonathan”, b) that those two individuals were equally likely to be a doctor, or c) that Jonathan is more likely to be a doctor than Elizabeth. A chi-square test suggested that responses on this measure differed by information condition,  $\chi^2(2) = 9.61$ ,  $p = .008$ , with more openness to Elizabeth being the doctor in the counter-stereotypical condition (2.8%) than the stereotypical condition (0.3%), and more openness to the two individuals being equally likely to be the doctor in the counter-stereotypical condition (25.7%) than the stereotypical condition (21.6%). There was no effect of materials,  $\chi^2(2) = 2.19$ ,  $p = .335$ . These results were not meaningfully different in an ordinal regression, which also demonstrated no interaction between materials and information, Wald(1) = .551,  $p = .458$ . Adding the combined standardized implicit measure to the ordinal regression revealed a significant positive relationship, estimate = .268,  $SE = .102$ , Wald(1) = 6.88,  $p = .009$ , suggesting that higher implicit judgments of Jonathan (vs. Elizabeth) as doctor-like (vs. nurse-like) predicted greater likelihood of expecting someone named Jonathan to be a doctor vs. equal

chances of either person being a doctor, and greater likelihood of either person being a doctor vs. Elizabeth being a doctor. Implicit judgments remained significant, estimate = .307,  $SE = .105$ , Wald(1) = 8.585,  $p = .003$ , when controlling for explicit judgments, which were independently significant, estimate = .169,  $SE = .081$ , Wald(1) = 4.39,  $p = .036$ .

## **Discussion**

The results of Study 12 highlight a variety of important considerations regarding the generalization of robust implicit revision to real-world applications. Though the work presented in Chapter II found strong reversals of implicit impressions when participants learned extreme new information that proved their initial assumptions to be erroneous, the findings of Cao and Banaji (2016) that counter-stereotypical clarifications of the professions of two individuals did not reverse implicit gender stereotypes of those individuals raises the question of whether new information needs to be extreme in order to prompt revision. Can less extreme but still seemingly diagnostic information that one was wrong about a person not be similarly effective?

The results of the present study suggest that the answer may be “yes”, and point to the nature of the measure and the relevance of information as potentially important moderators. While replicating Cao and Banaji’s (2016) basic effect (of no counter-stereotypic implicit impression formation) with a modified IAT, Study 12 found significant counter-stereotypical impressions of professions on a misattribution task (the PMP). Furthermore, counter-stereotypical impression formation was stronger when the individuating information was presented in an immersive, self-relevant narrative than when using Cao and Banaji’s (2016) original materials, finding that the immersive version reduced implicit gender stereotyping to being indistinguishable from neutral. In a combined, standardized form that facilitated a comparison of the effects of the information on each measure, the effects of individuating

information and materials (immersive vs. minimal) did not differ between the measures; this suggests that although the modified IAT continued to show higher mean levels of implicit gender stereotyping than the PMP, the two measures were equally good at reflecting the encoding of the individuating information and the deeper impact of the immersive story.

At present, the proper interpretation of the measure effect can only be a matter of speculation. A strong candidate explanation, as mentioned in the introduction, is that the IAT puts participants into a “categorical” frame of mind (De Houwer, 2001) by imposing a categorization task, which makes the gender categories of Jonathan and Elizabeth more salient, allowing gender stereotypes to continue to impact responses. Relatedly, more general gender stereotypes could have continued to be activated on the IAT by the novel nicknames for Jonathan and Elizabeth that were presented as stimuli on the task (e.g., jon, John, johnny, ell, Ella, eliza), which had not been learned and practice by participants as referring specifically to this Jonathan and this Elizabeth; for these reasons, the IAT used in this study may not have been completely altered to specifically measure implicit impressions of the particular Jonathan and Elizabeth in question, despite using their images as category *labels*.

Another possibility is that the difference stems from other dissimilarities between the tasks; for example, the IAT requires fast responses and scores on the task are computed via averaging of response times, whereas the PMP does not emphasize rapid responses and scores are calculated only based on the final judgments made on each trial. Automaticity features cannot be collapsed into a single construct (Bargh, 1994; De Houwer & Moors, 2012; De Houwer, Teige-Mocigemba, Spruyt, & Moors, 2009; Moors & De Houwer, 2006), and although both tasks are indirect measures of unintentional judgments of Jonathan and Elizabeth, variation in the extent to which responses on the AMP and IAT exhibit other automaticity features (like

speed), and differences in their operative processes (e.g., response interference vs. misattribution; De Houwer, 2001; Payne & Lundberg, 2014) make it likely that different implicit measures will sometimes produce different results (e.g., Deutsch & Gawronski, 2009). In the present case, for example, it is possible that the processing of the gendered names on some trials fleetingly activates general gender categories and corresponding stereotypes in the moments before they are individuated as applying to the particular Jonathan or particular Elizabeth depicted in the study. As a rapid-response measure, the IAT may allow for this fleeting activation to drive responding; on the slower PMP, on the other hand, the primes may be individuated before an impression of the pictograph coheres, allowing for the individuated impression to be misattributed to the target rather than a gender stereotype. This possibility should be investigated by future work.

Equally important are the implications of the immersion manipulation, which showed that even on the IAT, a more self-relevant way of delivering counter-stereotypical information about Jonathan and Elizabeth heightened the effectiveness of that information in attenuating stereotype-based implicit impressions. This may indicate that in a broad sense, even mundane revelations that we might have been wrong about a person may be capable of updating our implicit responses to that person to fall more in line with what we now know to be true about him or her. It also raises an important direction for future research: Determining why implicit responses might need more immersive information to reverse than explicit responses.

Finally, the results of Study 12 demonstrated that the implicit impressions of the professions of Jonathan and Elizabeth predicted a variety of other judgments that participants made about them. Though these results were exploratory and thus warrant a priori replication, the most interesting findings were arguably the correlations between the implicit impressions and

the more general beliefs the participants held about gender and the medical profession, because these relationships persisted even when controlling for explicit beliefs. These relationships found the implicit impressions to independently predict the views of the participants on the likelihood of men and women to be doctors, and their sense of the views of the public on those matters more generally. The prediction of these various measures by the implicit impressions of the characters adds support for their validity as measures of impactful responses. It also suggests, however, that future research should test whether the relationship between the implicit impressions of Jonathan vs. Elizabeth and participant beliefs about general gender stereotypes is due to a continued impact of gender stereotypic knowledge on implicit impressions of the two individuals (regardless of the individuating information they received), or due to the use of implicit impressions about Jonathan and Elizabeth to inform responses on those subsequent measures of societal notions of gender roles.

### **General Discussion**

The studies in this chapter tested the generalizability of implicit revision through reinterpretation, approaching that endeavor in several different ways. Collectively, these investigations into more varied evidence for reinterpretation as a route of implicit change not only speak to the broader external validity of the results obtained in Chapter II to implicit updating in diverse contexts, but in doing so, provide opportunities to identify and begin to test moderators and mediators of the updating process.

Studies 7 and 8 found that other approaches to implicit updating which share similar constituent features with reinterpretation, drawn from recent theory – negation of some element of earlier learning, coupled with affirmation of a new, countervailing impression – could be equally effective in overturning an initial implicit impression. These studies thus open the door



to a broader family of routes of implicit revision that simultaneously implement some form of both of these constituent processing steps, which collectively may be more effective at quickly revising implicit evaluations than negation or affirmation alone (Gawronski et al., 2008; Peters & Gawronski, 2011; Rydell & McConnell, 2006).

Study 9 tested a different, yet critical, element of the external validity of reinterpretation as a route of updating by addressing whether reinterpretation could revise implicit evaluations after a delay of days that could have produced deleterious forgetting of specific details of the story necessary for effective revision. Instead, reinterpretation was still highly effective in reversing implicit evaluations, regardless of recall – suggesting, perhaps, that construal of the new information at the time of its learning was the more important factor in determining its impact.

Next, evidence for reinterpretation as a mechanism for updating implicit impressions was found even in the context of a controversial societal issue (big-game hunting; Study 10) and in the face of implicit anti-Black bias (Study 11). Finally, more complex evidence for updating in the face of much more mundane and less extreme information – that a man and a woman held counter-stereotypic jobs – emerged as well (Study 12).

Though Studies 10-12 all found evidence for implicit updating, the strength of revision varied substantially among them. Together with the findings of Studies 7-9, this variation lends itself to some potential unifying principles on the generality of implicit updating.

Probably the clearest trend among the studies is that when the new information was both extreme and indisputable in its ability to compel participants to view the individual in a new light, revision was at its strongest, which might be expected. This was the case both with the Francis West fire rescue reinterpretation (Studies 7-9 and 11) and in the negate + affirm

condition (Studies 7-8), where the final information left little room for varied interpretation, and was arguably extreme; in the negate + affirm case, the positive detail that replaced the negated earlier events was both rare and made indisputably relevant by the accompanying negation. In the face of extreme and clear-cut information of this sort, revision may be likely to occur even against the backdrop of more general intergroup or visual-based implicit biases like race: Study 11 found robust revision (albeit slightly reduced) even when the character was Black, and even when implicit race bias occurred with the control faces. Furthermore, the results of Study 9 suggest that these ingredients may allow for successful revision even after a longer delay from initial formation, and even when recall of the specifics of the formative events is already fading.

What about when one or both elements – extremity and indisputable diagnostic value – are missing? Revision may be reduced, or at least more nuanced, in such cases; The information about big-game hunting in Study 10 may have been new and extreme (participants may have been surprised to learn that killing animals could be the best way to save them), but its diagnostic relevance to the individual hunter in that study may have been seen as less convincing and clear, and only one among many factors worth taking into account when considering him. To the extent that participants did report reinterpreting the activity, their implicit impressions of the hunter were updated, but in this study such reinterpretation may have been more of a motivated choice than an unavoidable consequence of merely learning the new information. The findings of Study 11 on counter-stereotypical professions, on the other hand, show that when the new details are seemingly clear and diagnostic but not particularly extreme (in that it is not uncommon to find women as doctors), implicit updating may be more nuanced, showing up on a misattribution task (AMP variant) but not a different implicit measure (IAT), and becoming stronger as the immersive nature of the new learning is increased. While the source of the disagreement between

measures remains speculative (and may ultimately be linked to systematic differences in the processes or automaticity features evoked by the tasks; Deutsch & Gawronski, 2009; De Houwer et al., 2009, 2012), it is sensible to remain cautious about whether apparently diagnostic but more minimal and mundane revelations will consistently produce updating.

Together, the considerations above raise the possibility that factors like how much time has passed since initial formation, and even whether the new information relates directly to the old information at all, may have their influence on implicit updating through their impact on the appraised extremity and diagnostic relevance of the new information encountered about a person. These factors, in turn, may be influential because of their ability to identify, reactivate, and update the relevant memory representations that contribute to the initial implicit response. In reviewing research from a variety of subfields of clinical psychology under a reconsolidation framework, Lane and colleagues (2015) argued that very different traditions (from psychodynamic approaches to cognitive-behavioral to humanist therapy) may all employ reconsolidation by helping patients to identify and modify currently maladaptive memory representations; psychodynamic probing of earlier history to find the root causes of current cognitive patterns may not be essential for successful change, but may provide a route for identifying currently problematic thoughts and responses and, by highlighting their irrational or no-longer-relevant origins, help patients appreciate the diagnostic value of the replacement habits of thought offered by the therapist; CBT, on the other hand, approaches this identification-and-change process more directly by attempting to reproduce maladaptive thought patterns (regardless of their origins) so as to identify and change them. When successful, the different traditions seem to all identify and reactivate maladaptive cognitions and then offer relevant replacements, which closely tracks the reconsolidation approach of reactivating a memory so as

to allow relevant new information to be integrated (Hardt, Einarsson, & Nader, 2010; Lane et al., 2005; Lee, 2009; Lee, Nader, & Schiller, in press).

An important direction for future research will be to test whether some of the factors offered by the Studies in Chapter III as potential moderators of revision are influential because of their ability to draw upon mechanisms of memory updating. For example, a longer time delay than that used in Study 9 might find revision to be less effective than after two days; if so, the study could explore whether this is due to the time delay preventing the new information from reactivating earlier memories of Francis. Perhaps after such a delay, the new information will no longer sufficiently reactivate even prior gist impressions of him, and more work would be necessary to retrieve the relevant memory traces to allow the new details to be integrated; this would fit with reconsolidation studies suggesting that memory age moderates the likelihood of updating (Alberini, 2007). Work on this topic is ongoing, but one contributing factor to this moderation may be that older memories are less likely to be reactivated during new learning (Gershman, Monfils, Norman, & Niv, 2017). It is also possible that at longer delays, the reinterpretation information would be more successful at revision than the negate + affirm condition, to the extent that the former does a better job of retrieving the representations that underlie the initial implicit response. Likewise, the individuating nature of the new information about the Black version of Frank in Study 10 may have prevented it from reactivating the set of representations that form the substrates of implicit race bias, preventing generalization of Frank's heroic actions to the Black control faces. A failure to reactive relevant causal memory traces and integrate new information into those traces may also have contributed to the persistence of gender stereotyping on the IAT even after learning counter-stereotypical information: Perhaps because such professions are not all that surprising or unheard of, and the information was so

minimal, participants may have failed to form strong or well-integrated representations of these novel individuals, resulting in category-based interference of stereotypes on a rapid response task that encourages categorical thinking (the IAT) even while the individuating information did emerge on a misattribution task (the PMP).

In sum, the findings of Chapter III offer many directions for future research, while increasing confidence in the potential for reinterpretation to play a wider role in implicit updating. One of the most important directions will be to attempt to generalize these mechanisms to the modification of group-based implicit impressions themselves (rather than just individual members of groups) – an issue to which I will return in the final General Discussion (Chapter V). In the next chapter, I turn from the question of external validity to a closer characterization of the automaticity features of the revised impressions.

## Chapter IV. Operating Characteristics of Reinterpretation-Based Revision

### I. Does Reinterpretation Lead to Revision that is Implicit?

Implicit measures are designed to tap responses that are unintentional (Ferguson, 2007; Ferguson & Fukukura, 2012; Payne et al., 2008). Research has suggested, however, that explicit response strategies can sometimes impact implicit measures when participants are motivated (such as through experimenter instruction) and have a strategy for doing so (e.g., Fiedler & Bluemke, 2005, Teige-Mocigemba & Klauer, 2013; see also Teige-Mocigemba & Klauer, 2008; cf. Degner, 2009).

A final goal of the present investigation was to test whether reinterpretation leads to the revision of truly *implicit* (unintentional) evaluations, and to more broadly explore the automaticity features of the revised impressions (De Houwer et al., 2009; De Houwer & Moors, 2012; Moors & De Houwer, 2006). This issue is of critical importance, because I am focused on the ways in which implicit evaluations can be revised. There are a couple reasons to question, however, whether recent findings are indeed demonstrations of change in implicit responses. First, the existing findings suggest that explicit evaluations in these studies tend to form and change in the same directions as implicit evaluations (though their correlations with other measures differ; see Chapter II, Table 1). If the implicit measures were to be “contaminated” by any explicit evaluation of the targets, the pattern in the implicit data could bear such general similarities to the explicit data even in the absence of truly implicit updating. Of course, the pattern is also consistent with genuine implicit and explicit updating.

Second, and more importantly, a great deal of the evidence in the present work for implicit revision has relied on the Affective Misattribution Procedure (AMP; Payne et al., 2005), a measure that has been challenged recently on its implicitness (Bar-Anan & Nosek, 2012).

Specifically, Bar-Anan and Nosek (2012) provided evidence that AMP effects were driven in large part by a small subset of participants who reported that they had intentionally evaluated the primes. They found that when such participants were excluded, the good psychometric properties of the AMP (its reliability and the size of the priming effect) were greatly diminished. They argued from such data that the AMP may not be a good measure of unintentional impressions. On the other hand, Payne and colleagues (2013) argued that such reports might be post hoc inferences stemming from awareness on the part of participants, when asked to reflect on the task, of correspondence between the primes and their responses. Across three studies, they demonstrated: 1) that participants with larger effect sizes on the AMP were more likely to endorse either that they had intentionally evaluated the primes or that they had been unintentionally influenced by them, depending on which question was asked; 2) that the standard (implicit) AMP predicted behavior differently from an explicit version; and 3) that participants rarely reported being influenced by the primes when queried on a trial-by-trial basis.

Previous work on the construct validity of the AMP has thus done much to establish that the responses it measures vis-à-vis the primes are implicit (Gawronski & Ye, 2014, 2015; Payne et al., 2008; Payne et al., 2013). This work collectively suggests that AMP effects are driven by misattribution (Gawronski & Ye, 2014; Payne & Lundberg, 2014), and that such misattribution can occur even when participants are distracted from the features of the primes under consideration and unaware of any influence of the primes on their impressions of the targets on a trial-by-trial basis (Payne et al., 2013). Finally, the AMP shows expected dissociation from explicit measures, such as in being less susceptible to self-presentational motivations (Payne et al., 2008; replicated by Open Science Collaboration, 2015). This body of evidence strongly supports the status of the AMP as an implicit measure.

However, given that the present set of studies claims to provide novel evidence for faster revision of implicit impressions than some prior theories have posited, the operating characteristics of the responses measured by the AMPs used in these studies is of particular importance, and warrants independent empirical investigation. If the revision in prior reinterpretation work is relatively more explicit than currently argued and assumed, it may recast the theoretical meaning of the findings, making it less clear the degree to which they show evidence of implicit revision. If, on the other hand, revision from reinterpretation is implicit, the findings contribute to the theoretical literature on revision, and also speak to the practical possibilities of reversing implicit first impressions. In this Chapter, I directly test the implicitness of revision from reinterpretation.

The impetus for undertaking an independent empirical test of the automaticity features of the AMP was a novel, anomalous observation that I recently detected with my colleagues in the distributions of AMP data in the earlier studies reported in this package, described in the next section.

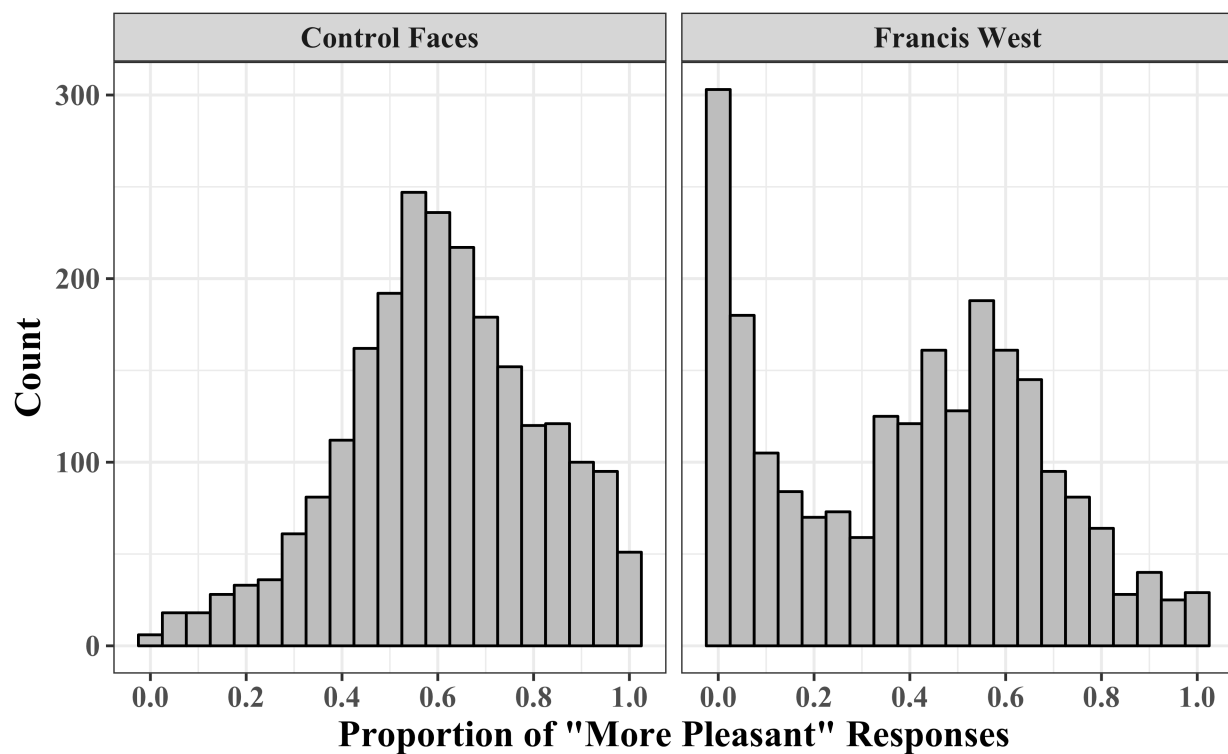
## **II. Novel observation of anomalous distributions in prior studies**

The critique raised by Bar-Anan and Nosek (2012), in suggesting that AMP effects may be driven largely by a subset of participants who intentionally rated the primes rather than the pictograph targets, focused on their finding that the high reliability and robust priming effects of the task were dependent on the subset of participants who agreed after the task that they had at least sometimes intentionally rated the primes (see discussion above of rebuttal by Payne et al., 2013). They did *not* discuss any evidence of unusual distributions in the AMP data.

In a survey of prior work reported in this package, however, an unusual and persistent pattern was detected in the frequency distributions of the AMP results. In the aggregate, the



pattern is most stark: Figure 19 shows how, pooling across Studies 1a and 2-9 (which feature the Francis West paradigm in its original form) the proportion of pictographs judged to be more pleasant than average at Time 1 (before the final information about Francis West) has a zero-inflated, bimodal distribution on trials in which Francis West is the prime stimulus, but not on trials in which a control face is the prime stimulus.<sup>18</sup> In other words, after learning that Francis West had broken into his neighbors' homes and caused destruction, many participants judged all (i.e., 20 out of 20 trials) pictographs on trials with Francis West as the prime to be less pleasant than average.

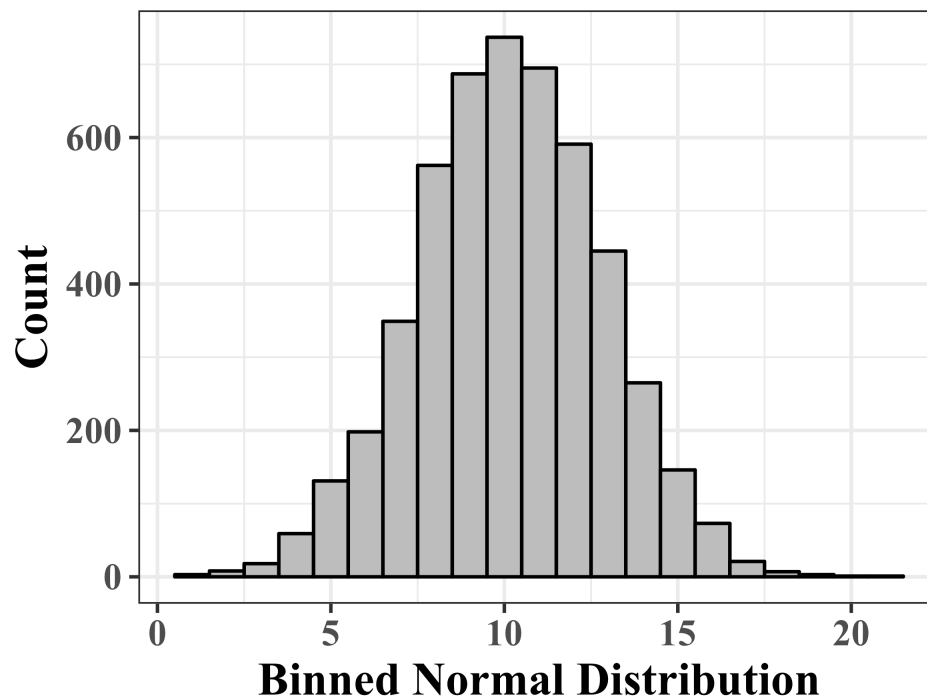


<sup>18</sup> Though Study 11 featured a version of the Francis West paradigm, it was substantively different from the other 9 studies in that there were different types of control trials and a new set of faces. For this reason, it will be discussed separately, and not included in the pooled analysis reported here.

*Figure 19.* Frequency distributions of the proportion of pictographs judged to be more pleasant than average at Time 1 by prime stimulus, pooled across information conditions and experiments (Studies 1a and 2-9).

Though the distribution of pleasantness proportions on Francis West trials is visually quite bimodal, bimodality can also be quantified using statistics like Hartigan's dip test (Hartigan & Hartigan, 1985), which tests the size of the maximum deviation of the experimental distribution from a unimodal distribution that minimizes that maximum deviation, and which has been recommended in social cognition research for formal tests of bimodality (Freeman & Dale, 2012; Hehman, Stoller, & Freeman, 2014). The pooled pleasantness proportions on Francis West trials, as displayed on the right side of Figure 15, do indeed show significant evidence of bimodality using this test,  $D = .066$ ,  $p < .001$ , and non-normality, Shapiro-Wilk  $W = .942$ ,  $p < .001$ . However, importantly, this statistical test of bimodality is not appropriate for use in assessing bimodality in the present experiments, because the interval (rather than truly continuous) nature of the data produce false positives on the dip test. Because most of the studies in this investigation include 20 trials per prime type, pleasantness proportions for Francis West can only take on values between 0 and 1 in intervals of .05, and no values in-between. The accumulation of scores at these values tends to register as a deviation from unimodality *even if the distribution used to generate those values is normal*. To demonstrate this, I drew 5000 random values from a normal distribution and then grouped them into 21 equal-width bins (to simulate the 21 values that a proportion can take between 0 and 1 in increments of .05). These bins are presented in Figure 20. Although the underlying distribution of raw values does not significantly deviate from normality ( $W = .9997$ ,  $p = .714$ ) or unimodality ( $D = .003$ ,  $p = 1$ ), the

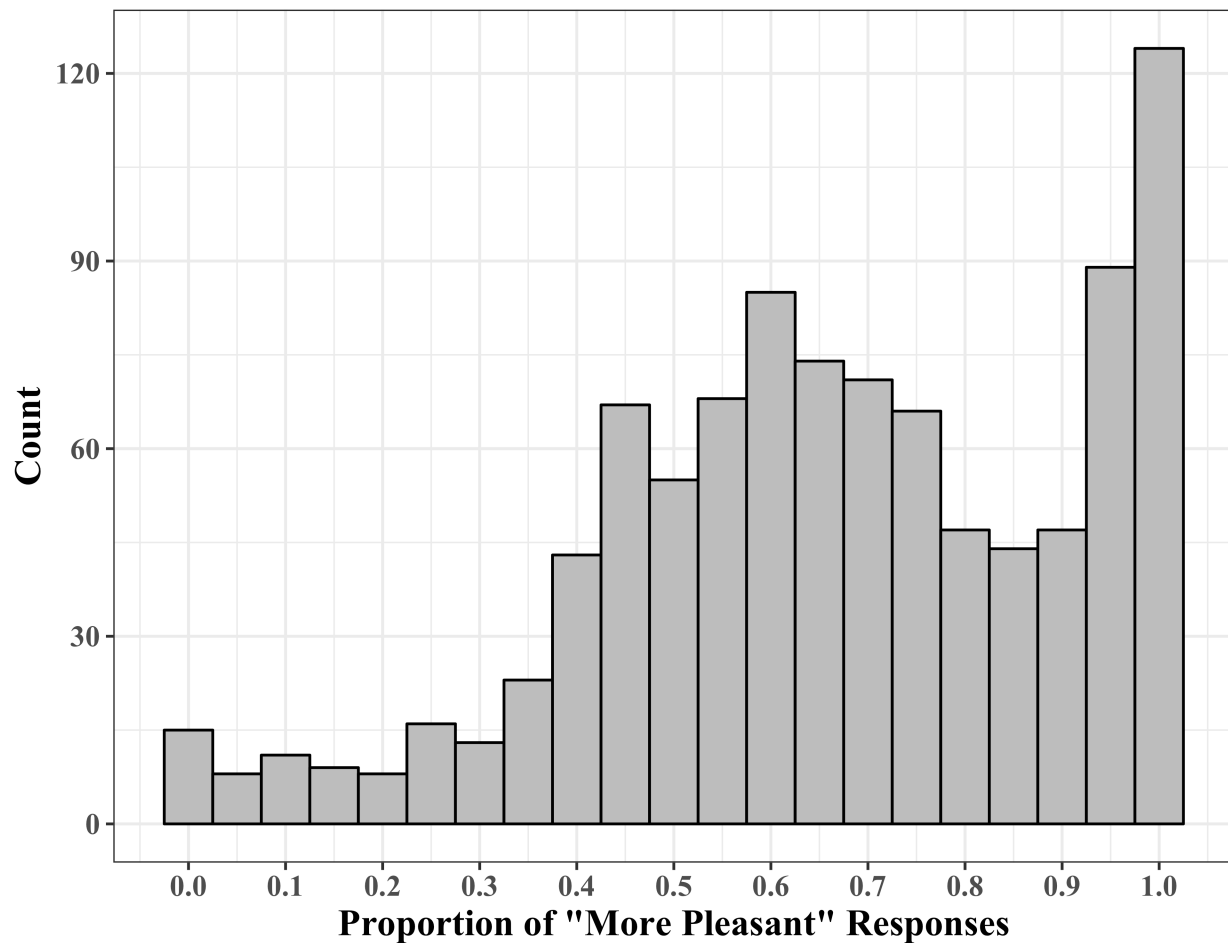
binned distribution is neither normal ( $W = .988, p < .001$ ) nor unimodal ( $D = .070, p < .001$ ) according to the tests. These tests, then, cannot be used to sort out whether a given AMP distribution is bimodal, which will be particularly problematic in subsequent studies in this chapter that attempt to eliminate the bimodality – the dip test cannot be used to determine if those efforts are successful. As such, simple visual inspection will be used, and frequency distributions will be regularly presented below for this purpose.



*Figure 20.* Values randomly drawn from a normal distribution split into 21 bins, which Hartigan’s dip test registers as multimodal.

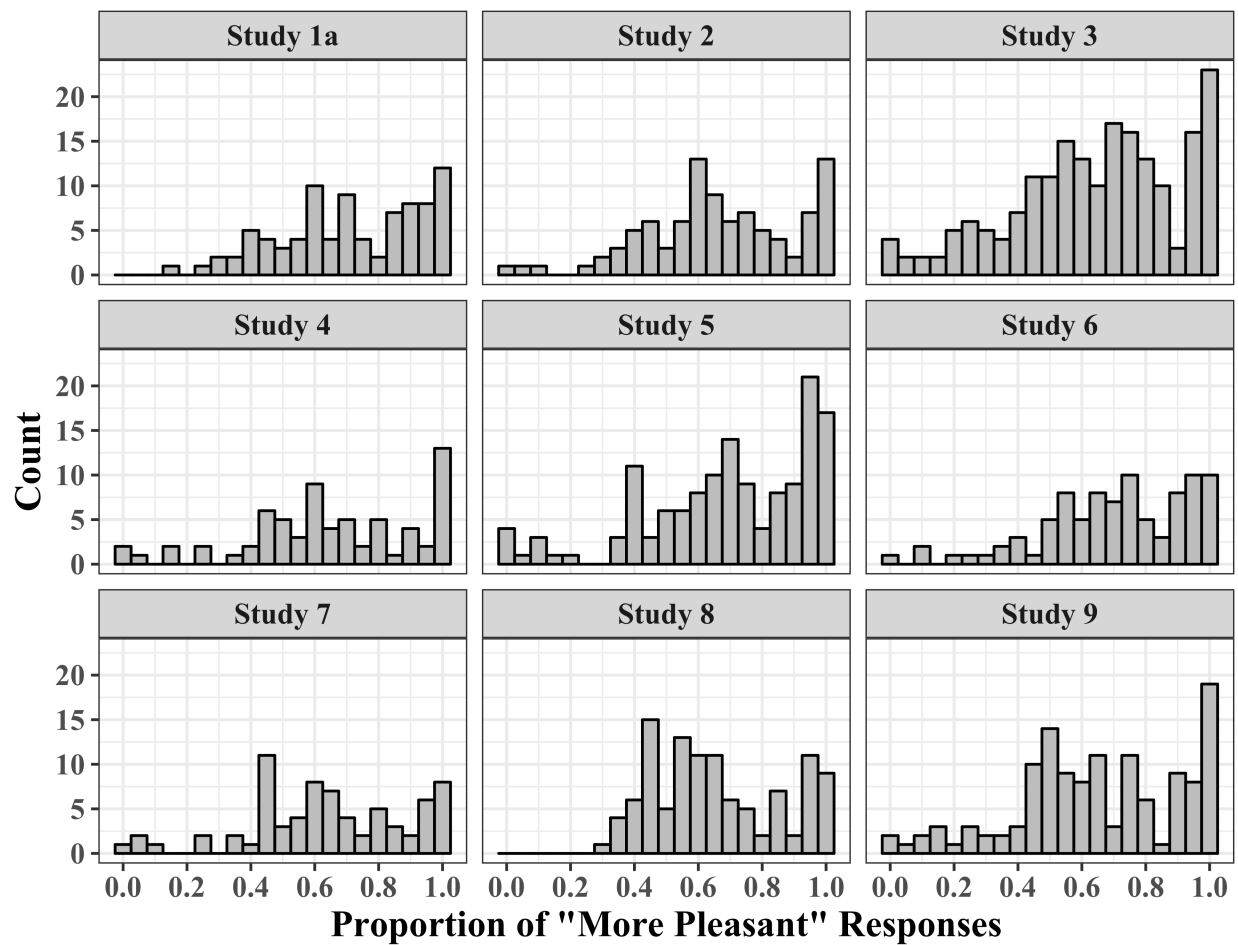
A similar, but opposite pattern of bimodality emerged at Time 2 (after the final information about Francis West) in the fire rescue condition; in that condition, across studies there emerged a second mode at ceiling (1), such that many participants judged all pictographs

on Francis West trials to be more pleasant than average after learning that Francis West was a heroic figure (Figure 21).



*Figure 21.* Frequency distribution of the proportion of pictographs judged to be more pleasant than average on trials in which Francis West is the prime stimulus at Time 2 in the Fire Rescue condition, pooled across experiments (Studies 1a and 2-9).

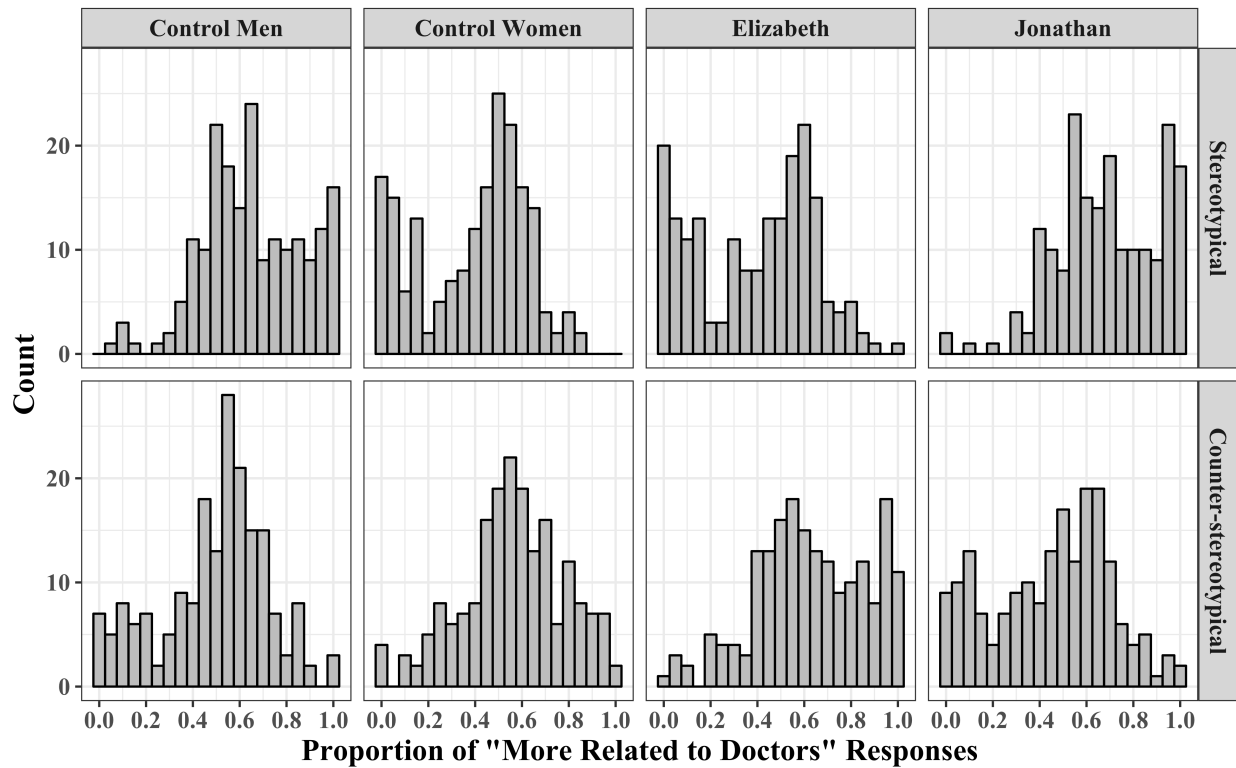
Some evidence for this bimodal pattern emerged not just in the aggregate, but within each study as well. Figure 22 shows the distributions of the frequencies in Studies 1a and 2-9 in the fire rescue condition at Time 2, after Francis has been revealed to be a hero.



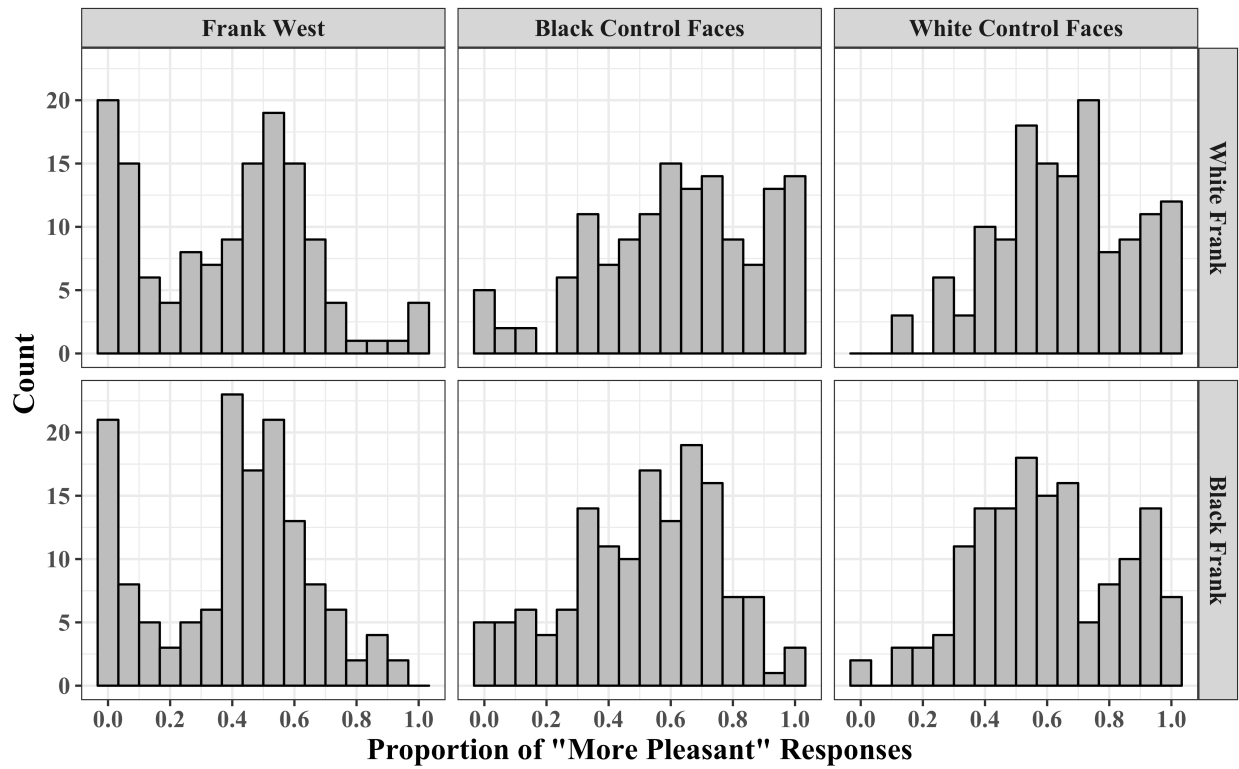
*Figure 22.* Frequency distributions of the proportion of pictographs judged to be more pleasant than average on trials in which Francis West is the prime stimulus at Time 2 in the Fire Rescue condition, Studies 1a and 2-9.

Likewise, there are collections of responses near floor (0) and ceiling (1) in Study 12, which examined stereotypical and counter-stereotypical implicit impressions of the professions of novel men and women (Figure 23) and similar quirks in the distributions of responses to Frank West in Study 11's examination of the interaction of individuating information and race at both Time 1 (Figure 24) and Time 2 (Figure 25). Frequency distributions did not appear idiosyncratic

in the examination of implicit impressions of a big-game hunter in Study 10, however (Figure 26).



*Figure 23.* Frequency distributions of the proportion of pictographs judged to be more related to doctors (vs. nurses) than average in Study 12, by prime stimulus and information condition.



*Figure 24.* Frequency distributions of the proportion of pictographs judged to be more pleasant than average at Time 1 in Study 11, by prime stimulus and race of Frank.

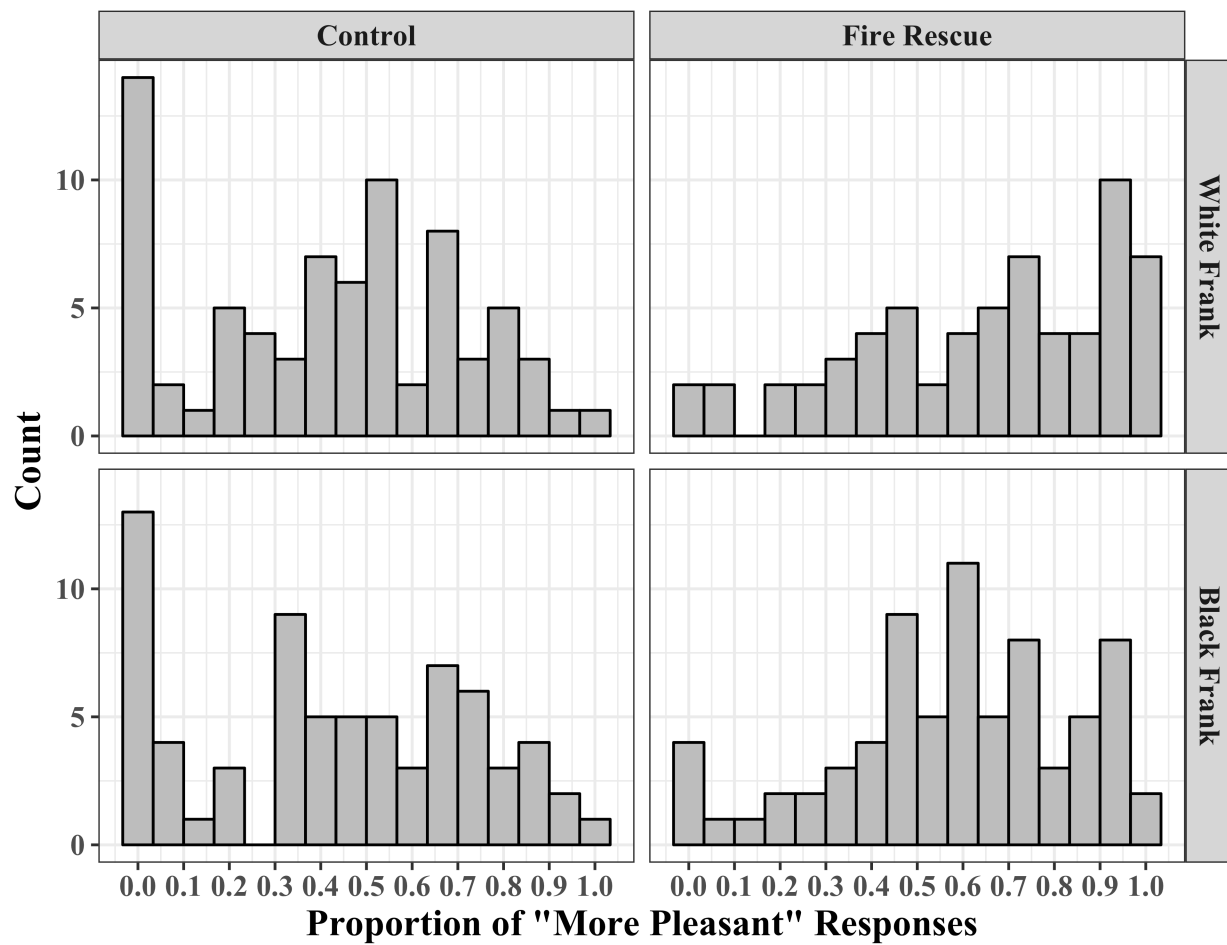


Figure 25. Frequency distributions of the proportion of pictographs judged to be more pleasant than average on trials in which Frank West is the prime stimulus at Time 2 in Study 11, by information condition and race of Frank.



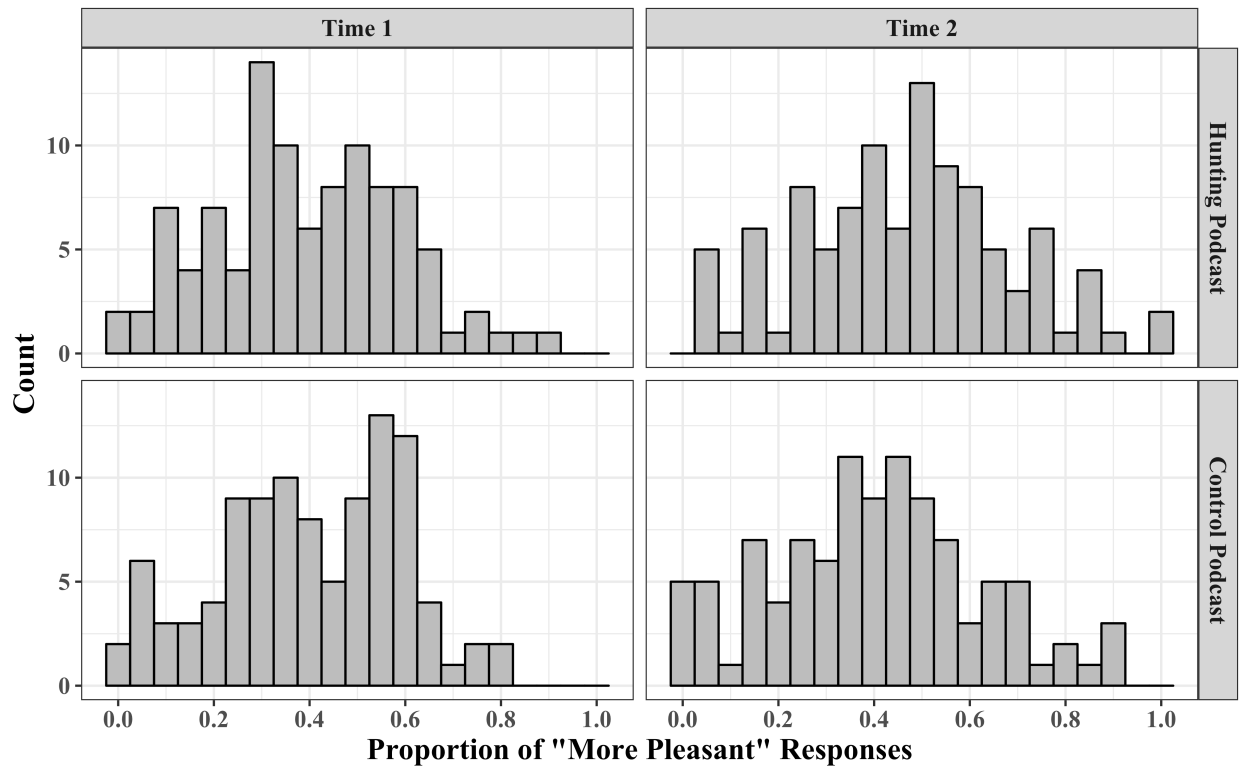


Figure 26. Frequency distributions of the proportion of pictographs judged to be more pleasant than average on trials in which Knowlton (the big-game hunter) is the prime stimulus in Study 10, by time and podcast condition.

The observation of bimodality in many of these distributions raises fresh concerns about whether the AMP effects presented as evidence of implicit updating in these lines of work truly provide evidence for shifts in *implicit* responses, or instead reflect some degree of explicit (intentional) judgments. The reason why this observation is potentially problematic is that bimodality in a distribution is often assumed to reflect the operation of distinct processes in the generation of the data (Freeman & Dale, 2012). In the case of zero-inflated distributions with count data, bimodality may occur when one process governs whether a case will be zero or not-zero (e.g., a decision about whether to participate in a game or not participate), and a separate

process governs the distribution of scores that are not-zero (e.g., how many points are scored by a person who does choose to play; see, e.g., Chester & DeWall, 2017). In aggregate group data such as what I presented above, the bimodality can also reflect two overlapping subgroups; for example, the raw distribution of a dependent variable may be bimodal after a strong manipulation (e.g., self-reported happiness in a sample of participants in which half received an electric shock and the other half received a \$10,000 check). Along these lines, in the case of the AMP distributions in the studies reported here, the bimodality could conceivably be produced through two processes: (1) Participants who choose to disregard the instructions and explicitly judge the primes will show floor effects when Francis is negative (Time 1) and ceiling effects when Francis is positive (Time 2 in the fire rescue condition), and (2) the group of participants who follow the instructions will show a more normal distribution of scores away from the floor and ceiling. The presence of both processes in the data would produce the bimodal pattern from two overlapping distributions. Only for the latter set of participants are the responses being measured toward the primes *implicit* under the usual criterion of unintentional (Ferguson, 2007; Ferguson & Fukukura, 2012; Payne et al., 2008).

This bimodal pattern need not *guarantee* two distinct generative processes, however, or necessitate that one generative process is intentional and another unintentional. It is possible that this distribution reflects an unusual pattern in the underlying responses that participants have to the materials, rather than that a subset of participants fail to follow the instructions on the task. At both Time 1 and Time 2, the AMP may be detecting that participants tend to fall into only partially overlapping groups; for example, at Time 1, participants may either have a very strong negative reaction to the story about Francis or a mild one (perhaps from reminding themselves that the study is fictional and the character fake, or due to suspecting a positive ulterior motive to

be revealed later), with few participants falling in-between these “camps.” At Time 2 in the fire condition, participants may be similarly divided into groups: some may have powerfully positive feelings toward him (and thus respond near ceiling), while others have a milder reaction – perhaps again due to the fictional nature of the story, or from being “hung up” on a detail about Francis carelessly stepping on the family cat in the course of saving the children. There may be relatively few people who have a view of Francis (or the impression targets in the other studies) that falls genuinely in-between. This explanation, of course, requires extra work to account for why bimodality is not observed on *explicit* evaluations (as it is not; floor effects prevail before the heroic detail, and ceiling effects after). It is possible that participants feel compelled to *express* an evaluation of Francis consistent with the clear suggestion of the story, even if their private feelings are more variable – with the latter being detected by the AMP, given its reduced susceptibility to self-expression motives (Payne et al., 2008).<sup>19</sup>

A related question pertains to why this pattern occurs in most (but not all) of the present studies but has not been reported in other work using the AMP. Explanations for the apparent novelty of this pattern among AMP studies fall into multiple camps as well; it could be that the pattern is present in other studies but has gone unnoticed, or unreported for lack of clear explanation. If it is *not* present in other work, this could be because a) the demand characteristics in these experiments – with relatively extreme information – are particularly pronounced relative to other studies, leading some number of participants to willfully disregard the instructions, or b)

---

<sup>19</sup> The IAT in Study 1b did not show a bimodal pattern, but differs from the AMP in a number of ways, including that it is a relative measure of implicit evaluations of Francis vs. control faces. A parallel distribution on the AMP would be the difference scores of the pleasantness proportions for Francis West and control faces; these distributions of difference scores are not as consistently bimodal.

underlying implicit impressions of the novel targets in this paradigm genuinely fall into this unusual distribution, but take more normal distributions in other lines of work.

These various interpretations of the bimodal distributions are, of course, only speculative. For this reason, the analyses and additional studies in Chapter IV will focus on providing empirical evidence for the implicitness of some of the key effects described in earlier sections (Chapters II-III). I begin with additional analyses of earlier studies, which aim to provide some initial evidence for whether the bimodal distributions are likely to be the product of intentional judgments of the primes, before moving to additional experiments.

### **III. Reanalysis of earlier studies**

#### **Are participants near floor/ceiling still affected by qualities of the pictographs?**

A unique feature of Study 8, which fit a multinomial model to aggregate AMP data within three information conditions (Payne et al., 2010), is that two sets of pictographs were used that had been preselected to be inherently more pleasant (pleasant set) or less pleasant (unpleasant set) than most other pictographs, based on pilot data. This was a requirement of the AMP multinomial model, but also affords an additional analytic opportunity: the participants with the most target-dependent extremity in their AMP responses to Francis West can be checked for evidence of any influence of the pictograph valence group on their responses. If these participants were intentionally rating the primes instead of the pictographs, then the (relatively subtle) variation in average pleasantness among the pictographs would not be expected to impact their responses. If such participants were following the instruction to attempt to rate the pictographs, however, and their large priming effects reflect genuine implicit impressions that are misattributed to the pictographs, then the qualities of the pictographs themselves should contribute to their responses.

Visual inspection of the frequency distributions in each information condition and time point (before – Time 1, and after – Time 2, the final information) suggested that excluding participants with proportions of pleasant responses on Francis West trials less than or equal to .1 at Time 1 or Time 2 in the control information condition, and greater than or equal to .9 at Time 2 in the fire rescue and negate + replace conditions, visually corrected the distributions. A reanalysis of the factorial ANOVA from that experiment – 2 (Time: Time 1, Time 2) x 2 (Prime Person: Francis West, Control Faces) x 2 (Pictograph Valence: Positive, Negative) x 3 (Information Condition: Control, Fire Rescue, Negate + Replace) – on *solely* the 88 participants<sup>20</sup> who would be excluded based on one of these criteria replicated the main effect of pictograph valence,  $F(1, 85) = 8.17, p = .005, \eta_p^2 = .088$ . This demonstrates that participants with extreme reactions to Francis West, thereby contributing to the bimodal patterns, still show an overall impact of the qualities of the pictographs on their responses. This makes it unlikely that they are exclusively evaluating *only* the primes.

### **Do revision effects emerge when excluding participants near floor/ceiling?**

Another approach to assessing the robustness of revision effects to the exclusion of participants who may potentially be responding to the primes in an intentional manner is to re-run analyses while dropping participants with extreme scores, in the bimodal tails. This strategy is problematic in that it will also discard data from participants who are performing the task as intended and show the strongest effects, and reduces power; it can thus be considered a particularly conservative test.

To follow this route while keeping statistical power as high as possible given the extreme nature of this strategy, I reexamined revision effects in the Francis West paradigm in a pooled

---

<sup>20</sup> Of these 88, 37 (42%) met the criteria at both Time 1 and Time 2, 31 (35%) met the criteria solely at Time 1, and 20 (23%) met the criteria solely at Time 2.

dataset of all participants included in Studies 1a and 2-9 while modeling the effects of experiment (Time 2 Fire Rescue distributions shown previously in Figure 22). Specifically, I ran a mixed linear model using the SPSS linear mixed procedure, including only participants with Francis West pleasantness proportions between .1 and .9 (exclusive). Fixed factors included Time, Prime Type, Information Condition, and all of their interactions. To examine the robustness of the most critical comparison between the fire rescue (reinterpretation) and subway rescue (heroic but non-reinterpretation) conditions, only these information conditions were included. Time and Prime Type were repeated measures with an unstructured covariance matrix. A random effect captured variation across experiments in the critical 3-way interaction between time, prime type, and information condition.

Even on this subset of participants, the 3-way interaction was significant, indicating differential revision between the fire rescue and subway rescue conditions,  $F(1,33.18) = 6.59, p = .015$ . Simple effects tests showed that Francis West was more negative than control faces at Time 1 in both information conditions (both  $ps < .001$ ); at Time 2, Francis West remained significantly less positive ( $M = .52, SD = .17$ ) than control faces in the subway condition ( $M = .58, SD = .19$ ),  $F(1,37.94) = 11.58, p = .002$ , but Francis West had become significantly more positive ( $M = .58, SD = .16$ ) than control faces ( $M = .53, SD = .16$ ) in the fire rescue condition,  $F(1,27.23) = 35.10, p < .001$ . There was no evidence of significant variance in the effect attributable to experiment, Wald  $Z = .716, p = .474$ .<sup>21</sup> This shows that the Francis West revision effect does not depend on participants in the extreme tails of the distributions, and that the nature

---

<sup>21</sup> All of these results are equivalent if Experiment is treated as a fixed factor fully crossed with the other three and/or if only the three experiments that included both the fire rescue and subway rescue conditions are included (to equate cell sizes), all  $ps < .05$ .

of the new information (whether or not it prompts reinterpretation) continues to critically matter even when setting aside participants with extreme scores on the AMP.

**Do participants near floor/ceiling differ in other identifiable ways?**

If participants who have extreme scores toward Francis West on the AMP (low proportions of “pleasant” trials when Francis is negative and high proportions when he is positive) are systematically different from other participants in their strategy for approaching the task – such as by being more likely to have chosen to deliberately judge the primes instead of the targets – then it might be possible to detect other differences between those with extreme scores and those with less extreme scores. After all, if these participants are not following the task instructions, then their reasons for doing so could also lead to differences on other measures. Were such participants confused? Were they less careful in reading the materials (including instructions) presented during the study? Explicit measures included in the various experiments provide opportunities to test for clues. For example, participants in the extremes of the distributions – if they are “intentional responders” – might report that they found the task more confusing, or might score worse on comprehension checks if they were less attentive overall. If participants with more extreme scores are not so different from other participants in their approach to the experiment, on the other hand, then differences on other measures may be less likely (though could still occur to the degree that those differences are causally related to variance in implicit impressions, rendering findings of significant differences inconclusive).

To look for differences between participants with extreme scores (falling into the bimodal tails) and the rest of the sample with high power, I pooled all participants from the 7 experiments dealing with the basic Francis West effect discussed above that included the AMP as well as explicit questionnaire items of potential interest (Studies 1a, 2-6, and 9). Based on examinations

of distributions across multiple studies, participants were classified as contributing to bimodality if their proportion of pleasant responses on Francis West trials at Time 1 was less than or equal to .1, if their proportion at Time 2 in the control information condition was less than or equal to .1, if their proportion at Time 2 in the fire rescue condition was greater than or equal to .9, or if their proportion at Time 2 in the subway rescue condition was less than or equal to .1 or greater than or equal to .9 (as this condition had signs of additional modes at floor *and* ceiling). Point-biserial correlations were calculated between this bimodality indicator and 12 self-report measures: confusion with the study (all studies), extent of thinking about the story (Studies 1a and 2), difficulty making sense of the story (Studies 1a and 2), perfect comprehension performance (all studies), change in meaning of the earlier story (Studies 4-5), whether thoughts came together quickly vs. gradually (Study 4), deliberation extent (Studies 4-5), belief that the story was real (Studies 6 and 9), positive mood (studies 6 and 9), recall score (Study 9), extent of thinking back to the new information after relieving cognitive load (Study 3) and extent of thinking back to the old information after relieving of cognitive load (Study 3).

Only two correlations were significant. Participants in the bimodal tails were more likely to be in a good mood at the end of the experiment,  $r(446) = .11, p = .019$ , and reported more extensively thinking back to the new information after relieving cognitive load in Study 3,  $r(246) = .128, p = .044$ . The interpretation of the former relationship seems unclear, and the latter would seem to suggest that people with more extreme scores were being *more* attentive to the procedures. No other correlation involving the bimodality indicator was significant, all  $ps > .16$ .

### **Can the distributions be modeled in terms of component AMP processes?**

As noted briefly in an earlier section, though the bimodal distributions are unusual, and could be consistent with a subset of participants responding intentionally to the primes, it is



possible that the bimodality stems from the distribution of underlying evaluations of the primes, even if no intentional responding is present. The AMP process model (presented in Study 8) offers one way to conceptualize the different processes that underlie any given response on the AMP that might contribute to the distribution of overall scores on the task. The M parameter models the likelihood that whatever response the participant has to the prime will be expressed toward the pictograph, while the A parameter captures the likelihood that the prime will activate a positive response, and finally the P parameter indicates the likelihood that the response to the pictograph will be pleasant in the absence of misattribution. Intentional responding would most likely entail a high value for the M parameter, as the likelihood that impressions of the prime will guide judgments of the pictograph are high in that case. The alternative that the distribution of reactions to the primes are not normally distributed would, on the other hand, likely be evidenced by a bimodal distribution of the A parameter. Additionally, pooling at the extremes of the distributions of AMP pleasantness proportions might be more likely to occur if M and A are correlated; this seems plausible in that a participant with a stronger reaction to the Francis West prime may have increased likelihood of misattribution, even if they are attempting to only judge the pictograph targets.

To examine these possibilities, I reanalyzed the results of Study 8 to fit an AMP multinomial model to the data from each participant *separately*. The main analysis in Study 8 fit a single model to the aggregate data from all participants within each information condition, to be consistent with common practice (Payne et al., 2010; see also Conrey et al., 2005; Sherman et al., 2008) and because models fit on individual participant data can be high in noise. Nonetheless, the distributions of the resulting fit parameters could add insight in addressing the underlying sources of the bimodality in the AMP distributions.

For each participant, there are 8 trial cells, from crossing the levels of Time (2), Prime Type (2), and Pictograph Valence (2), allowing for a maximum of 7 model parameters. I tested two alternative models:

*Model 1.*

4 A parameters (one per prime type per time, across pictograph valence)

1 M parameter (across pictograph valence, time, and prime type)

2 P parameters (one per pictograph valence, across time and prime type)

*Model 2.*

3 A parameters (one for control faces across time, and two for Francis West, at Time 1 and 2, across pictograph valence)

2 M parameters (one per time, across prime type and pictograph valence)

2 P parameters (one per pictograph valence, across time and prime type)

Model 2 was not able to reproduce the experimentally observed patterns in the data, so Model 1 was selected (Palminteri, Wyart, & Koechlin, in press). Model 1 generated predicted data that reproduced all experimentally observed patterns in the AMP data in Study 8 (Figure 27), and also reproduced the patterns of bimodality in predicted Time 1 data (Figure 28) and predicted Time 2 data (Figure 29).

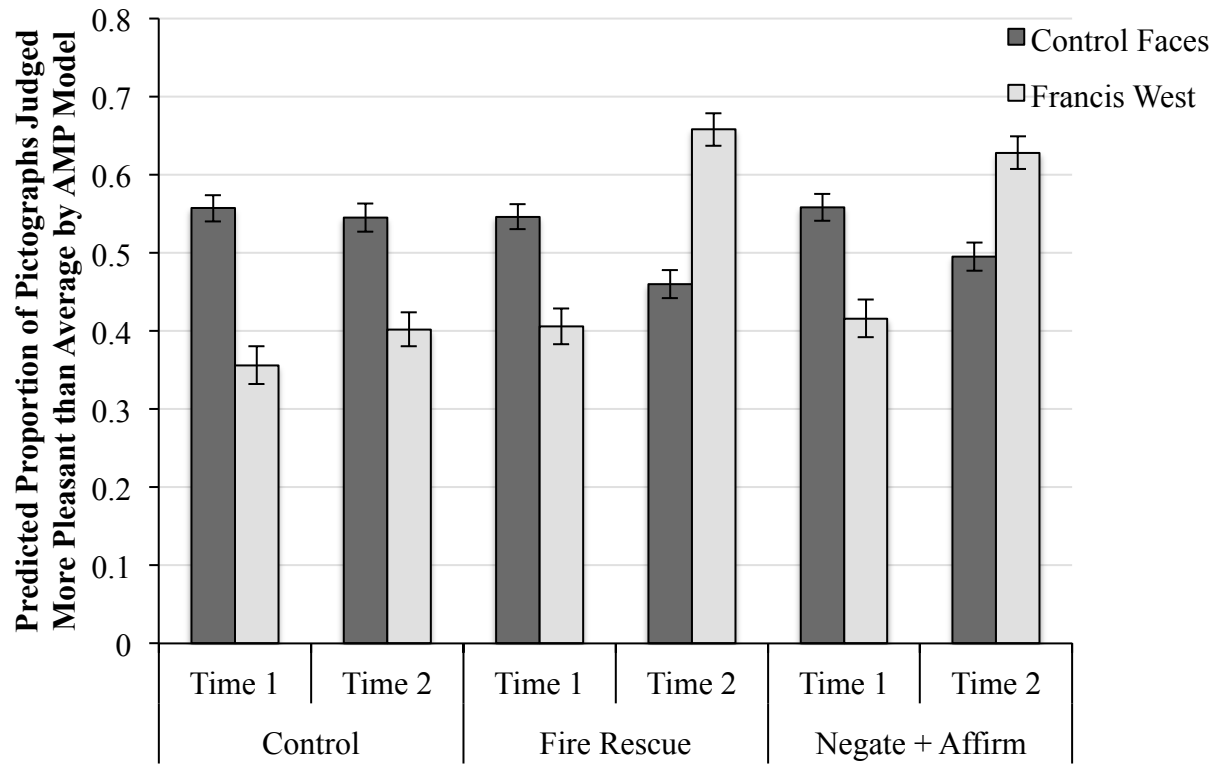


Figure 27. Predicted AMP results from multinomial model fit to individual participant data, which reproduce the results of Study 8. Error bars are standard errors.

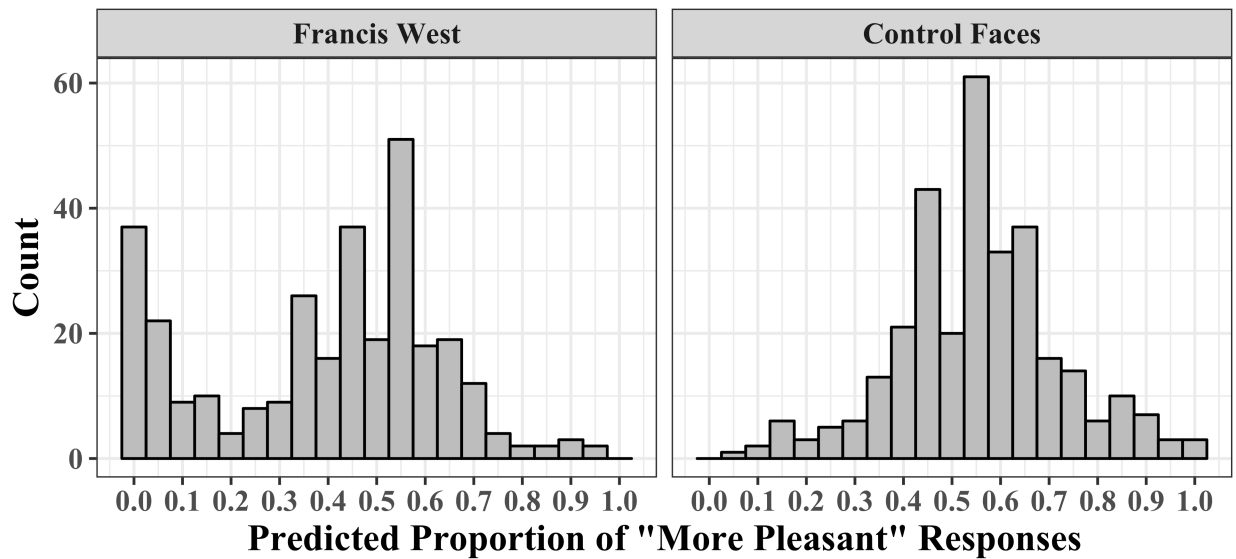


Figure 28. Frequency distributions of the predicted proportions of pictographs judged to be more pleasant than average at Time 1 by the AMP model, by prime type.

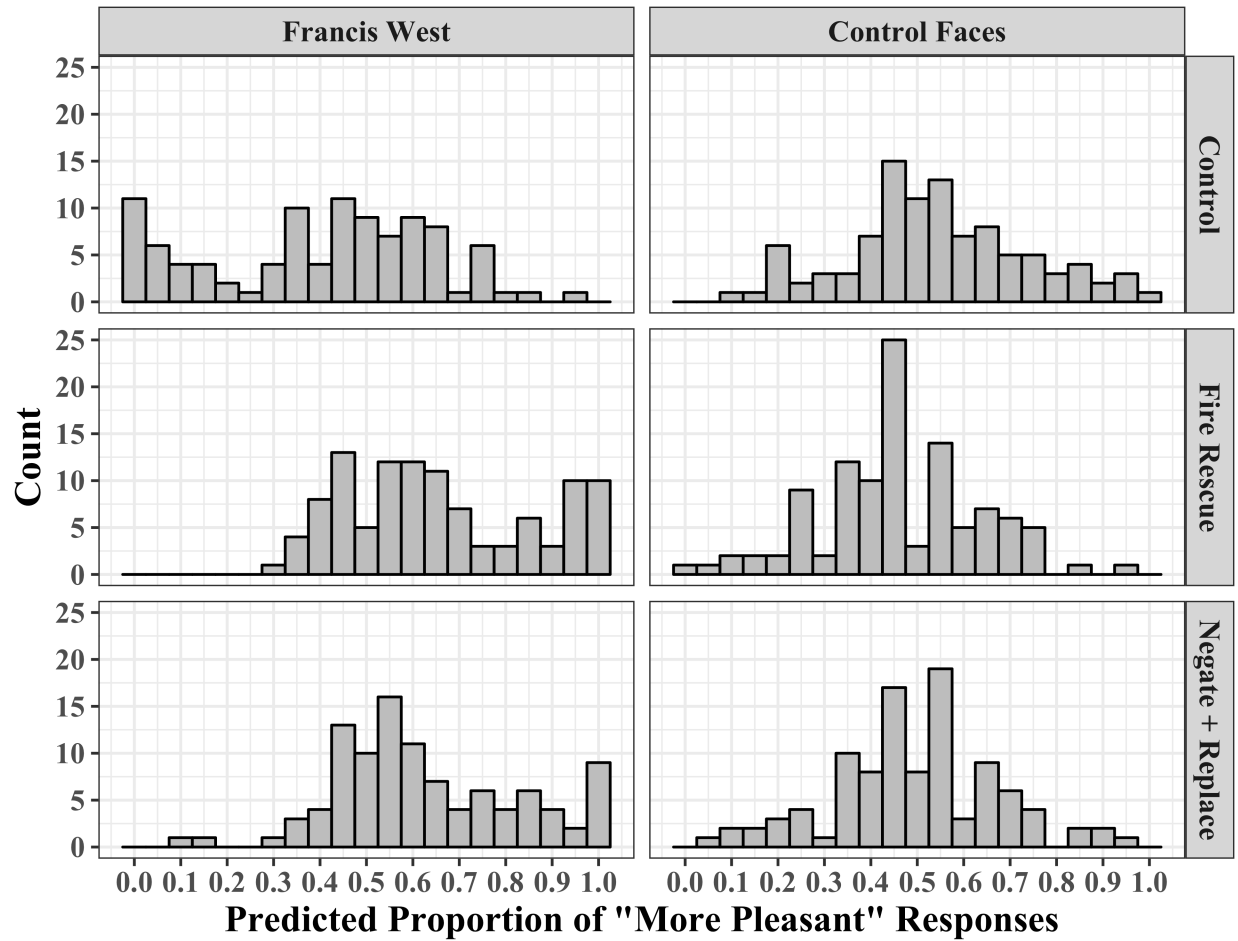
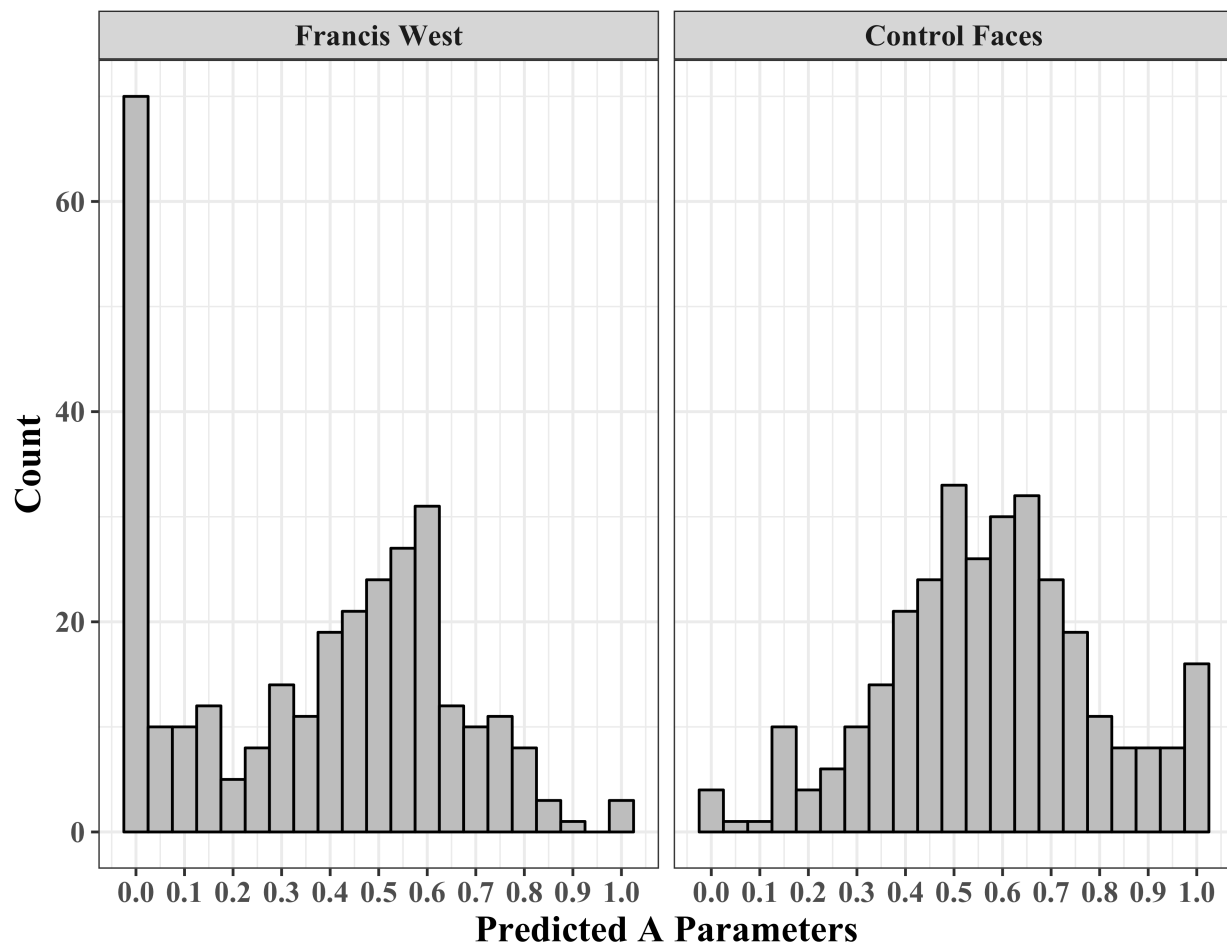


Figure 29. Frequency distributions of the predicted proportions of pictographs judged to be more pleasant than average at Time 2 by the AMP model, by prime type and information condition.

**Analysis of fitted model parameters.** Across the sample of 310 participants, the participant-level models had adequate fit (indicated by a non-significant  $p$ -value), with an average  $p$ -value of .167 ( $SD = .164$ ). The distributions of A parameters were markedly bimodal, both at Time 1 (Figure 30) and Time 2 (Figure 31). The M parameter was bimodal in the control information condition and more weakly so in the other two (Figure 32).



*Figure 30.* Frequency distributions of the A parameters fit by the AMP model at Time 1, by Prime Type.

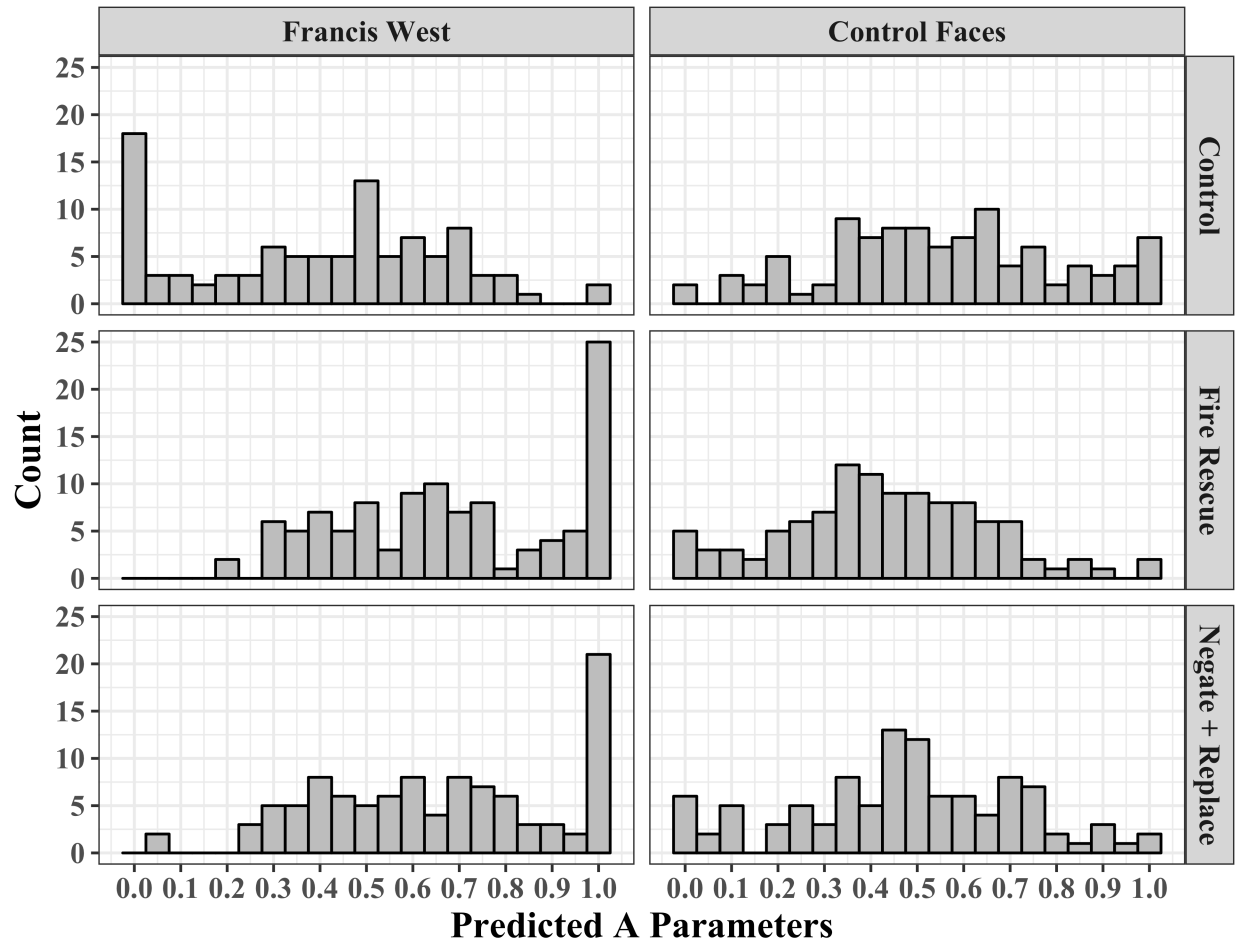


Figure 31. Frequency distributions of the A parameters fit by the AMP model at Time 2, by Prime Type and Information Condition.

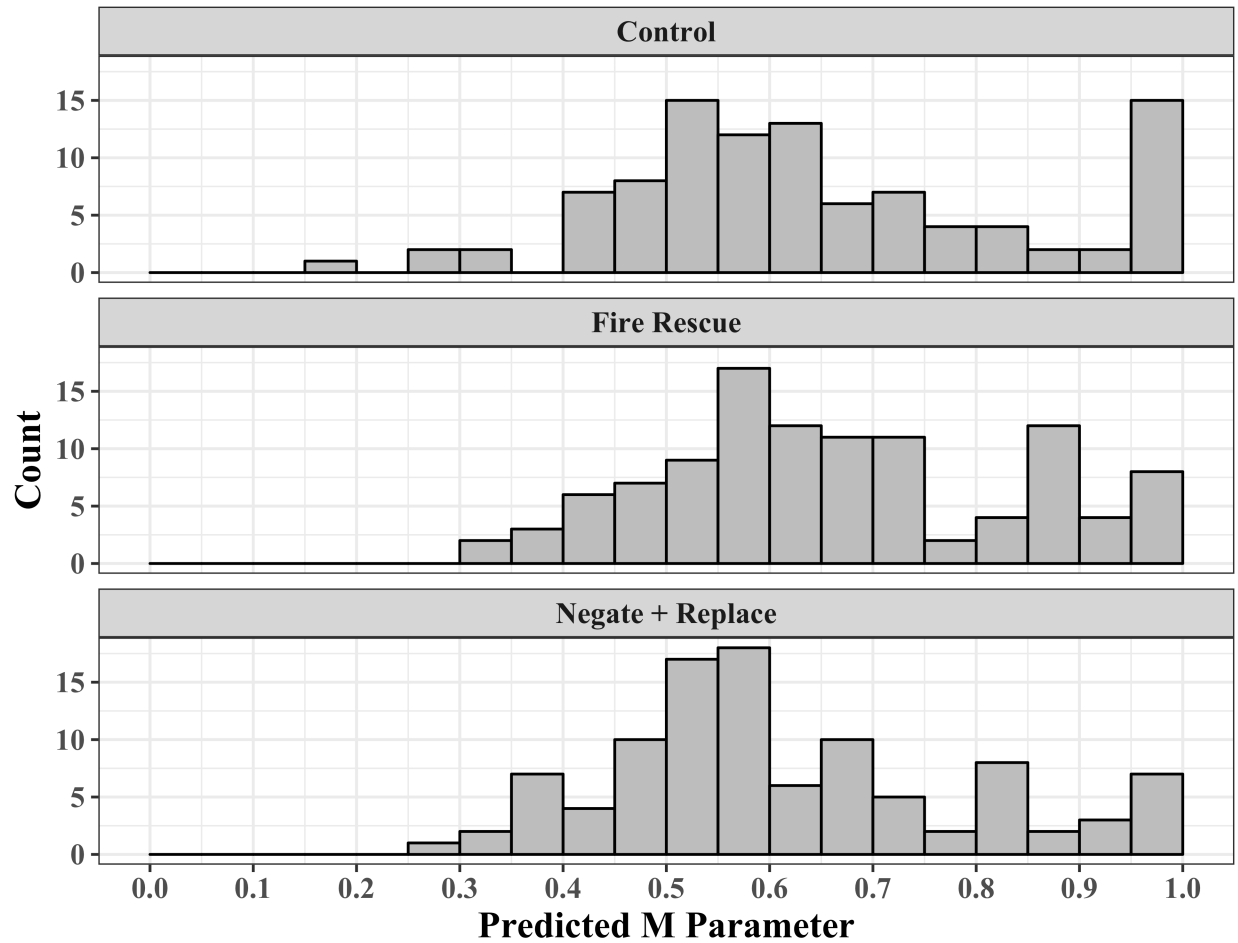


Figure 32. Frequency distributions of the M parameters fit by the AMP model, by Information Condition.

There were strong correlations between the M parameter and the A parameters corresponding to evaluations of Francis West, suggesting that participants with more extreme responses to the primes co-occurred with a higher likelihood of that response influencing ultimate judgments of the pictograph. At Time 1, there was a negative correlation between the M parameter and the A parameter for Francis West,  $r(308) = -.526, p < .001$  (but no relationship with the A parameter for the control faces,  $r(308) = -.042, p = .459$ ). At Time 2, in the control condition, the relationship between M and the A parameter for Francis West was similarly

negative,  $r(308) = -.462, p < .001$ . In the fire rescue condition, on the other hand, the relationship between M and the A parameter for Francis West was positive,  $r(308) = .575, p < .001$ ; the same was true in the negate + replace condition,  $r(308) = .365, p < .001$ . In sum, for participants with stronger negative reactions to Francis West when he was believed to be bad and more positive reactions to him when he was believed to be good, the likelihood of responding to the pictograph based on the prime was higher, compounding to produce many participants with AMP proportions near floor or ceiling.

These fitted parameters, which were able to reproduce all of the patterns in the data, suggest that bimodality in evaluative reactions to the primes is capable of producing the experimentally-observed AMP patterns even when the distribution of the likelihood of misattribution is less unusual overall, at least in the fire rescue and negate + replace conditions. Furthermore, a correlation between the intensity of evaluative reactions to the primes and the likelihood of misattributing those reactions to the pictographs – even when honestly attempting to judge only the pictographs – is an intuitively plausible pattern, even if no participant in the sample is intentionally judging the primes rather than the pictographs.

The additional analyses conducted in this section collectively argue against the possibility that the bimodal patterns observed in the AMP data are produced (at least solely) by a subset of participants who choose to disregard the task instructions and intentionally rate the prime images instead of the pictographs, and also demonstrate that the critical analyses are robust to the exclusion of participants with extreme scores falling into the tails of these distributions. However, additional empirical data is necessary to more firmly test the implicitness assumptions of the AMP as used in the current work. It is to these studies that I now turn.

#### **IV. Additional studies to test implicitness**



To bolster the case for the implicit nature of the revision effects found in earlier studies, a number of additional experiments will be presented that approach the question in different ways. Following one of the critical experiments reported by Payne and colleagues (2013; Study 3), Studies 13 and 14 will reproduce two central effects from earlier studies while offering participants the option to skip any trial during the AMP on which they felt their impending response to the pictograph might be influenced by the prime face image. To the extent that participants skip few trials and still show evidence of priming effects on the non-skipped trials, the argument that participants are intentionally rating the primes online during the task is counter-indicated. Such persistent effects speak not just to the unintentional nature of the measured responses to the primes, but also to lack of participant awareness of such influence. This is arguably better evidence for lack of intentional responding to the primes than post-hoc survey questions that are open to confabulation (Payne et al., 2013). Importantly, if bimodality is detected even when the skip option is offered, it supports the proposal that this pattern is caused by something other than intentional responding.

Study 15 will then make use of a technique proposed by Gawronski and Ye (2015) for decoupling post-hoc reports of intentional responding (whether accurate or confabulated) from priming effects on the AMP. This study will provide participants in the Francis West procedure with a distracting secondary task to perform while completing the AMP that is meant to draw their attention away from the identity of the prime faces (Francis West or unknown control faces), so as to assess whether the priming effects emerge in the same direction even when no longer related to self-reported intentional responding. Because of the taxing nature of the distraction, this study will also add further evidence of the effects of cognitive load on implicit revision (see Study 3).

Finally, Studies 16 and 17 will add an additional, forceful “warning” page to the instructions of the AMP to strongly discourage participants from intentionally rating the primes during the task. These studies examine the impact of this warning on the distributions of responses on the task as well as the prime effect within the control, fire rescue (reinterpretation), and subway rescue (non-reinterpretation positive action) information conditions (see Study 2), with a particular focus on whether the critical difference between the fire rescue and subway rescue conditions continues to emerge even with this warning.

### **Study 13: Awareness of Influence: An Option to Skip – I**

Bar-Anan and Nosek (2012) raised the possibility that on the AMP – the implicit measure used in most of the work on fast implicit impression revision, including studies in the present work (Cone, Mann, & Ferguson, in press) – a large portion of the priming effect could be driven by intentional evaluation of the primes, because effect sizes on the AMP correlated with self-reported extent of intentionally rating the primes (when it is the *pictographs*, and not the primes, that participants are instructed to judge). Payne and colleagues (2013) found, however, that such a correlation was probably based on some post hoc realization on the part of participants that there had been a systematic relationship between the primes and their responses through the course of the task. For instance, such participants were also simultaneously more likely to endorse the idea that they had been unintentionally influenced by the primes (suggesting that they might endorse any plausible theory connecting the primes and responses to which the researchers drew their attention), showed very different patterns of responding when they were explicitly instructed to rate the primes, and showed little awareness of being influenced by the primes at a more sensitive trial-by-trial level. These results indicate that overall, the AMP measured unintentional evaluations, and perhaps even that participants are unaware of the

influence of the primes on the pictographs on a trial-by-trial basis, because they generally did not avail themselves of an option to skip trials on which they felt the prime was influencing their feelings toward the pictograph, and having this option available did not reduce the priming effect. It may be that only in hindsight (when the correlation between their responses and the primes on the trials is clear in the aggregate) that inferences of intentional responding develop.

The fast revision work, however, is very different from that of Payne et al. (2013) in that it directly provides participants with strong reasons to change their mind about a target person that serves as the critical prime on the task, whereas the primes used in the work of Payne and his colleagues were members of established social groups about whom the participants did not learn new information during the course of the study. The information presented in the fast revision studies, which shows that a prior impression of an individual was highly inaccurate, could provide participants with a particularly strong incentive to exert explicit control over their responses on the AMP. This would suggest that “implicit” revision effects might actually reflect some degree of *explicit* responding, which would undercut the theoretical contribution of the reinterpretation findings. I expected, however, that reinterpretation changes the unintentional impression of the target person that is primed. These competing possibilities promote very different ways of viewing the efficacy of fast revision effects; do these manipulations only overturn the effects of a rejected first impression when participants are motivated to explicitly regulate their responses, or do they generalize to cases in which person impressions truly operate implicitly?

To test these different possibilities, I adapted a version of the study discussed above (Payne et al., 2013, Study 3) in which participants were provided with an option to skip any trial of the implicit measure when they felt that their response to the pictograph would be influenced

by the prime that came before it. I predicted that if positive reinterpretation of seemingly negative initial behaviors produced a revised implicit impression of the target, then I would continue to observe a priming effect comparable to when no option to skip trials is available. Alternatively, if explicit evaluation of the primes is the primary reason for a strong revision effect on the measure, then I should observe that the availability of the “skip” option reduces or eliminates the priming effect, because those participants who for whatever reason will not or cannot evaluate the pictographs and would otherwise evaluate the primes instead now have a legitimized way to proceed through the task without violating the instructions. Because my focus was on whether participants would intentionally evaluate the primes on the AMP after learning the fire rescue information, all participants were assigned to the reinterpretation (fire rescue) condition, and I administered the AMP only once in the current study, after that final information about Francis had been presented.

## **Method**

**Participants.** Fifty-nine students were recruited through a department subject pool in return for partial course credit ( $M_{\text{age}} = 19.61$  years,  $SD_{\text{age}} = 1.59$  years, 13 men, 46 women) and randomly assigned to either the *trial skip option* or *no trial skip option* condition.

**Learning task.** After completing a short unrelated study, participants read through the story about Francis West used in the prior two studies.

Participants were next told that they would now read some additional information about Francis West, and were all presented with the reinterpretation condition, learning that Francis had broken into the homes of his neighbors to save children from a fire. Participants were instructed to take at least 15 seconds to consider the new information, and then advanced to the next task at their own pace after at least that much time elapsed.

**Implicit measure.** After participants learned the information about Francis West, I assessed implicit evaluations of him using an AMP similar to those used in the preceding studies. The specific instructions for this task varied depending on trial skip condition.

***No trial skip option condition.*** In this condition, participants completed the AMP using the conventional instructions, as in Studies 1 and 2. The AMP included 60 trials (30 with Francis West as the prime, and 30 with control faces as the prime).

***Trial skip option condition.*** Participants in the trial skip option condition completed the AMP in an identical manner to those in the no-skip condition, with one exception. Drawing on a similar condition in prior work (Payne et al., 2013, Experiment 3), I instructed participants that they should skip a trial, by pressing the spacebar, if they felt that the prime on that trial might be influencing their response to the pictograph. Specifically, they read:

*“IMPORTANT: If you ever think that your evaluation of any pictograph might be influenced by the photo that preceded it, you should pass on the trial by pressing the space bar. You should only evaluate the pictograph whenever you believe that your opinion reflects the qualities of the pictograph itself. Otherwise, please skip the trial whenever you feel your evaluation of the pictograph is being influenced by the photo that came before.”*

Each trial of the task contained a reminder on the bottom of the screen that the spacebar should be used to skip the trial if they felt influenced by the photo that preceded the pictograph. In every other regard, this condition was identical to the standard no-skip condition.

Two sets of 60 pictographs drawn from prior work (Payne et al., 2005) were used in this study as target stimuli (regardless of skip option condition), with one set randomly assigned for each participant. Pictograph set had no effects in the analyses reported below.

**Explicit attitude scale.** After the AMP, participants completed the same six-item scale from the previous studies to assess their explicit attitude toward Francis West.

**Final questionnaire items.** Finally, participants were asked whether they know Mandarin and/or Cantonese, to identify Francis West from the set of 11 face images presented during the study, to identify the final information they read about Francis West from a set of five short summaries, to provide their age and gender, and to give open-ended feedback regarding what they thought the study was about and any other comments they had. They were then debriefed, thanked, and compensated.

## **Results**

**Data preprocessing.** In accordance with established protocol for the AMP (Payne et al., 2005), I excluded *a priori* 18 participants who speak Mandarin and/or Cantonese, as the pictographs would not be neutral for such participants. This left a final sample of 41 participants. However, because of the large number of exclusions due to language in this undergraduate sample (30.5%), I also chose to report (unplanned) analyses involving results for the full set of 59 participants. All participants correctly identified Francis West from the set of 11 faces at the end of the study, and all but one participant correctly identified the final information they had read about Francis West.

**Proportion of skipping on Francis vs. control trials.** Overall, participants in the skip option condition opted to skip trials relatively infrequently. In the restricted sample, participants skipped an average of 3.70 trials out of 60; in the full sample, they skipped an average of 3.89. Even so, if participants have any on-line subjective awareness of being influenced by the primes (whether intentionally or unintentionally), then they would nonetheless be expected to be more likely to skip trials with a Francis West prime than trials with a control prime, because the former

have stronger evaluative connotations than the latter. If, on the other hand, the misattribution account is correct, and participants attribute an evaluative response triggered by the Francis West prime as being generated by the pictograph itself, then participants may be equally likely to skip Francis West and control trials. They might, in fact, be more likely to skip control trials, in that those trials are more difficult given their lack of any subjective evaluative reaction to the targets (Payne et al., 2013).

The results suggest that participants lacked insight into being influenced by the Francis West prime, supporting the misattribution hypothesis. In the restricted sample, participants were directionally more likely to skip trials with control primes ( $M = 2.40$  trials out of 30,  $SD = 4.03$ ) than trials in which Francis West was the prime image ( $M = 1.30$  trials out of 30,  $SD = 2.96$ ),  $t(19) = 1.59, p = .128$ , Hedges'  $g_{av} = .30$ . In the full sample, this difference was significant (control primes:  $M = 2.57$  trials,  $SD = 3.66$ ; Francis West prime:  $M = 1.32$  trials,  $SD = 2.70$ ),  $t(27) = 2.423, p = .022$ , Hedges'  $g_{av} = .38$ .

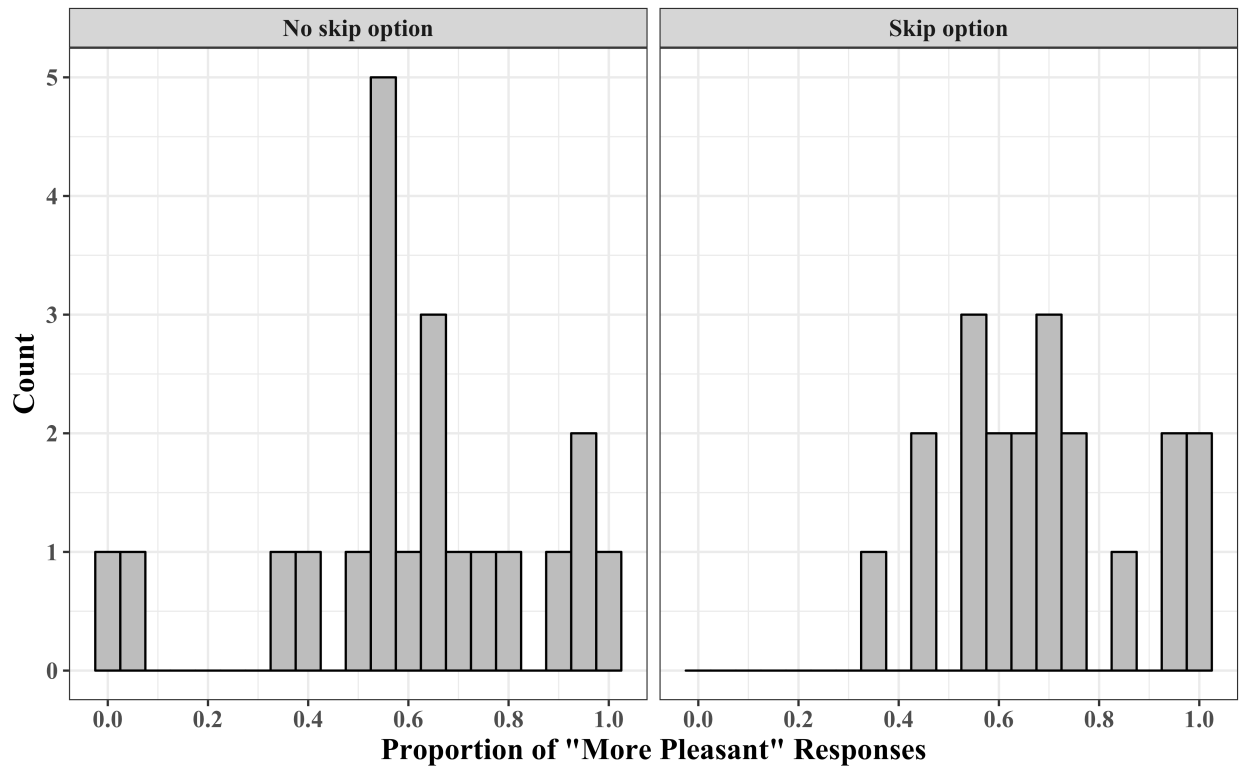
**Implicit evaluations.** I next compared implicit evaluations of Francis West (relative to control faces) in the skip and no-skip conditions. If participants lack insight into how the primes are influencing their responses to the targets, then having the option to skip should not decrease the size of the priming effect. For each participant, I calculated the proportion of trials on which the pictograph was judged to be more pleasant than average following the Francis West prime image, and a similar proportion for trials following control face images. For participants in the skip option condition, these proportions were calculated using the number of non-skipped trials.

I analyzed these proportions in a 2 (Prime: Francis West, control faces) x 2 (Skip Condition: trial skip option, no trial skip option) mixed ANOVA, with the latter factor manipulated between-participants. There was a main effect of prime in both the restricted

sample,  $F(1,39) = 14.296, p = .001, \eta_p^2 = .268$ , and full sample,  $F(1,57) = 11.296, p = .001, \eta_p^2 = .165$ , such that pictographs were judged as pleasant more frequently on trials with Francis as the prime (restricted sample:  $M = .647, SD = .230$ ; full sample:  $M = .632, SD = .228$ ) than on trials with control primes (restricted sample:  $M = .449, SD = .189$ ; full sample:  $M = .487, SD = .209$ ). There was no main effect of skip condition in the restricted sample,  $F(1,39) = 2.087, p = .157, \eta_p^2 = .051$ , or the full sample,  $F(1,57) = .182, p = .671, \eta_p^2 = .003$ . The effect of prime was not moderated by skip condition in either the restricted sample,  $F(1,39) = .306, p = .584, \eta_p^2 = .008$ , or the full sample,  $F(1,57) = .165, p = .686, \eta_p^2 = .003$ . Simple effects tests found that the priming effect was significant in both the no-skip condition (restricted sample:  $F[1,39] = 5.341, p = .026, \eta_p^2 = .120$ ; full sample:  $F[1,57] = 4.599, p = .036, \eta_p^2 = .075$ ), and the skip-condition (restricted sample:  $F[1,39] = 9.167, p = .004, \eta_p^2 = .190$ ; full sample:  $F[1,57] = 6.753, p = .012, \eta_p^2 = .106$ ). Collectively, these findings suggest that the priming effect was not impacted by the availability of the skip option, suggesting that participants were not aware of, and could not exert control over, the misattribution of their evaluative responses from the primes to the targets.

**Bimodality.** As shown in Figure 33, despite the availability of the option to skip any trial in which the participant felt that their response to the target might be affected by the prime, the distribution of the proportion of pleasant responses on trials in which Francis West was the prime was still bimodal.





*Figure 33.* Frequency distribution of the proportion of pictographs judged to be more pleasant than average on trials in which Francis West was the prime stimulus in Study 13, by skip option availability. Proportions were computed out of the total number of non-skipped trials.

**Explicit evaluations.** Though the primary focus was on implicit evaluations, I also tested for the effect of skip option condition on explicit liking of Francis West. The six items in the liking scale cohered well (Cronbach's  $\alpha = .834$  in the restricted sample, .870 with the full sample), and were thus averaged together to create a single score for explicit liking. Skip option condition had no effect on explicit liking in either the restricted sample,  $t(39) = .372, p = .712$ , Cohen's  $d = .116$ , or the full sample,  $t(57) = 1.565, p = .123$ , Cohen's  $d = .410$ . This suggests that having the skip option available on the AMP did not subsequently impact how participants deliberately evaluated Francis West. Instead, he was explicitly liked on average, and

significantly so (above the scale midpoint of 4) in the restricted sample ( $M = 5.89$ ,  $SD = .74$ ),  $t(40) = 16.32$ ,  $p < .001$ , and in the full sample ( $M = 5.79$ ,  $SD = .85$ ),  $t(58) = 16.15$ ,  $p < .001$ .

## Discussion

In this study, participants who were encouraged to skip any trial on which they felt that their impression of the pictograph was biased by the prime only rarely invoked that option, were more likely to do so on control trials than on those featuring the target of the impression as a prime, and showed directionally *larger* priming effects compared to participants who were not provided with an option to skip trials (in fact, in the full sample this trend was significant). This suggests that participants were largely unaware of being influenced by the primes on a trial-by-trial basis, which would preclude their ability to exert intentional control over the priming effect. The results thus supported the hypothesis that reinterpretation would produce a positive impression of the target which could implicitly influence responding – that is, without intention.

In contrast, the results did not corroborate the alternative proposal that reinterpretation primarily affects explicit control over AMP responses, while leaving implicit impressions potentially unchanged. I note that although this study had a comparably small sample size, the sample was large enough to replicate the implicit preference for Francis West over control primes in the reinterpretation condition, even within the skip option and the no-skip conditions. Furthermore, a larger sample may well have found significantly more frequent skipping of control trials rather than Francis West trials, as I found in the full sample here, which would speak against the hypothesis that participants are aware of the influence of the Francis West prime on their impressions of the pictographs. Finally, I expected that if the AMP were truly driven by explicit responding in this paradigm, allowing participants to indicate this by skipping those trials would produce a rather large effect of the skip manipulation, which did not emerge.

Importantly, the present results complement the studies in Chapters II-III in that they inform a different aspect of how reinterpretation impacts the impressions measured by the AMP. Though earlier studies dealt with the process operations that occur as a result of reinterpretation – both at the time of learning (negation + affirmation), and during responding (shifts in impression activation but not misattribution) – the current evidence speaks to the automaticity *conditions* under which those processes occur (like intentionality, awareness, and control; see, e.g., Gawronski & Bodenhausen, 2014; Moors & De Houwer, 2006). By demonstrating that reinterpretation results in a positive impression of the target that can be impactful even when participants are able to skip any trial on which they believe the prime might be biasing their judgments of the pictograph, the results entail that the AMP is not capturing intentional responses to the primes; in fact, participants do not even seem to be aware of the prime effects while completing each trial.

The findings also add a point of clarity to the source of the bimodality observed in the distribution of the Francis West AMP pleasantness proportions. By replicating the bimodal pattern (at Time 2 in the fire rescue condition) even when participants were offered – and encouraged to use – an easy way to avoid allowing the prime to influence their judgments in *any* way that they had conscious access to (whether through intentional responding or otherwise), this study suggests that the bimodal pattern may have emerged through the implicit misattribution mechanism that research has generally pointed to as the driving force behind AMP effects (Payne & Lundberg, 2014). To gather more evidence to bolster this point, Study 14 offered participants a similar skip option on an AMP in the very different paradigm featured in Study 12, in which bimodal distributions were observed on a profession misattribution procedure (PMP; see Figure 23).

## Study 14: Awareness of Influence: An Option to Skip – II

### Method

**Participants.** Out of a targeted sample size of 400 (to achieve 200 participants per each between-participants condition), 398 participants completed the study. Two additional participants reported technical issues and were compensated for their participation, but generated no data. Out of the initial set of 398 participants, data from six participants were dropped for familiarity with Mandarin or Cantonese, and one additional participant was excluded for using a single key on every trial of the PMP. This left 391 participants for analyses (53.2% women, Age  $M = 36.67$  years,  $SD = 11.61$ ).

**Procedure.** Participants completed a procedure similar to that from Study 12, in which information about two novel individuals (Jonathan and Elizabeth) is presented that either conforms to gender stereotypes of typical professions of men and women (Jonathan is a doctor and Elizabeth is a nurse) or defies gender stereotypes (Elizabeth is a doctor and Jonathan is a nurse). The two chief differences between Study 12 and the present study were that the procedure was greatly simplified, and that the skip option used in Study 13 was made available to all participants.

All participants were placed into the minimal information condition. As in Study 12, they were first acquainted with Jonathan and Elizabeth and told that either individual might be the doctor or nurse. They were then assigned to the stereotypical or counter-stereotypical information condition, with the new information being presented on a single screen as done in the work by Cao and Banaji (2016), with the one exception being that the same faces used in Study 12 were again included. The more immersive information condition from Study 12, in which participants were asked to imagine going through a medical emergency, was not included

here. Furthermore, there was no manipulation of measure: All participants completed the PMP, rather than the IAT, so that the availability of an option to skip trials could be examined. As in the previous study (Study 13), the PMP was modified to provide participants with the option to skip any trial in which they felt that the prime stimulus might influence their response to the pictograph, by pressing the space bar. That option appeared on-screen on every trial, along with the standard key labels of “More related to doctors” and “More related to nurses.”

## Results

**Implicit impressions.** I assessed implicit impressions of the primes – Elizabeth, Jonathan, control women, and control men – as more associated with doctors vs. nurses by computing the proportion of *non-skipped* trials in which the pictograph was judged as more related to doctors (vs. nurses) within each prime type. These proportions were then analyzed in a 4 (Prime Type: Elizabeth, Jonathan, Control Women, Control Men) x 2 (Information Condition: Stereotypical, Counter-stereotypical) mixed ANOVA, with prime type manipulated within-participants. For this analysis, data for 3 participants were set aside, because they skipped every trial within at least one prime type, making the calculation of all 4 proportions impossible.

There was a significant main effect of prime type,  $F(3, 1158) = 20.55, p < .001, \eta_p^2 = .051$ . This was qualified by a significant interaction between prime type and information condition,  $F(3, 1158) = 51.88, p < .001, \eta_p^2 = .118$ .

An examination of the simple effects showed that in the stereotypical information condition, in which Jonathan was revealed to be the doctor and Elizabeth was revealed to be the nurse, Jonathan was more implicitly associated with doctors (vs. nurses;  $M = .66, SD = .22$ ) than Elizabeth ( $M = .40, SD = .23$ ),  $t(168) = 8.86, p < .001$ , Hedges’  $g_{av} = 1.13$ . Jonathan was also more associated with doctors than control men ( $M = .61, SD = .21$ ),  $t(168) = 3.20, p = .002$ ,

Hedges'  $g_{av} = .22$ . However, Elizabeth was no more associated with doctors (vs. nurses) than control women ( $M = .42$ ,  $SD = .22$ ),  $t(168) = 1.37$ ,  $p = .173$ , Hedges'  $g_{av} = .09$ . Control men were more implicitly associated with doctors than control women,  $t(168) = 7.10$ ,  $p < .001$ , Hedges'  $g_{av} = .912$ .

In the counter-stereotypical information condition, Elizabeth was more implicitly associated with doctors (vs. nurses;  $M = .60$ ,  $SD = .23$ ) than Jonathan ( $M = .49$ ,  $SD = .25$ ),  $t(218) = 4.11$ ,  $p < .001$ , Hedges'  $g_{av} = .46$ . Elizabeth was also more implicitly associated with doctors than control women ( $M = .51$ ,  $SD = .21$ ),  $t(218) = 5.15$ ,  $p < .001$ , Hedges'  $g_{av} = .41$ . Jonathan was less implicitly associated with doctors (thus, more implicitly associated with nurses) than control men ( $M = .52$ ,  $SD = .21$ ),  $t(218) = 2.16$ ,  $p = .032$ , Hedges'  $g_{av} = .15$ . Unlike in the stereotypical condition, control women and control men did not differ in their implicit association with doctors over nurses,  $t(218) = .65$ ,  $p = .515$ , Hedges'  $g_{av} = .07$ . Figure 34 shows the proportion of pictographs judged to be more related to doctors (over nurses) within each condition.

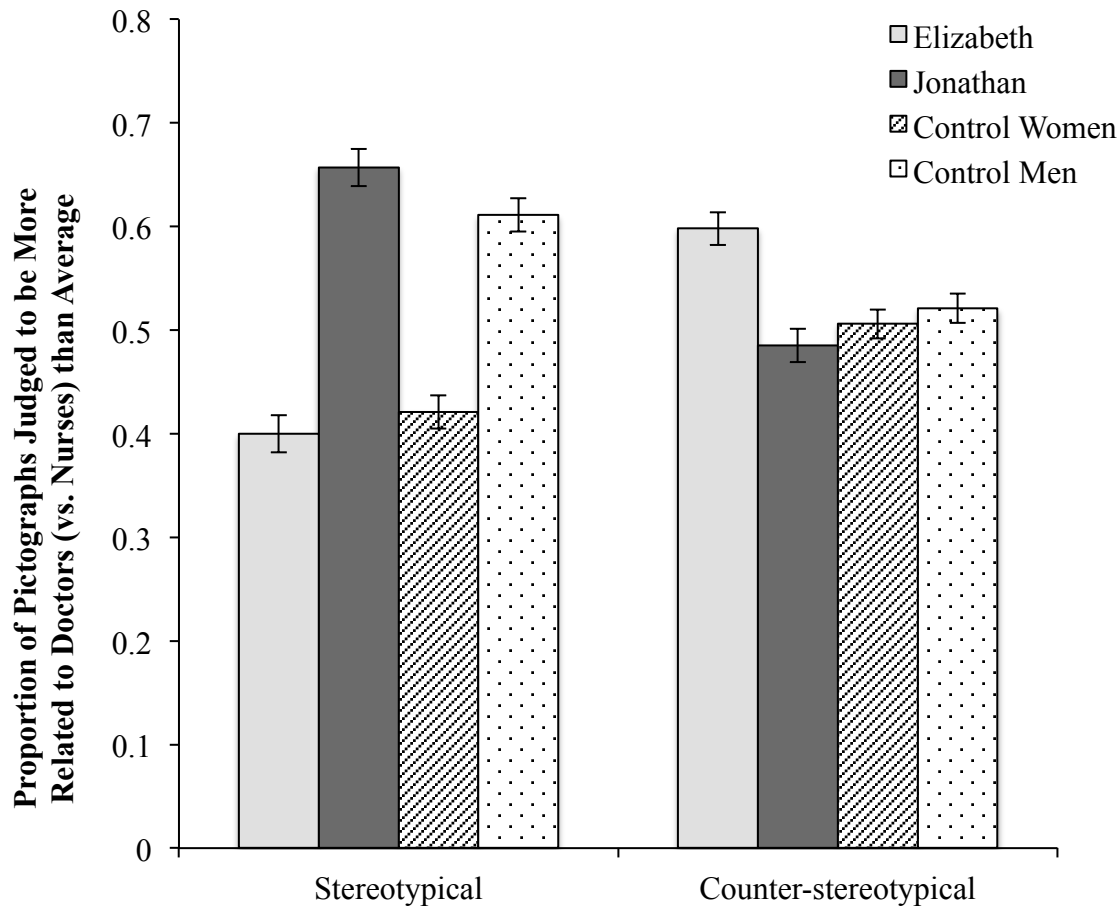


Figure 34. The proportion of pictographs judged to be more related to doctors (over nurses), by prime type and information condition in Study 14. Error bars are standard errors. Proportions were computed using non-skipped trials.

**Skip frequencies.** The proportion of trials that were skipped by each participant were computed within each prime type, and analyzed in a 4 (Prime Type: Elizabeth, Jonathan, Control Women, Control Men) x 2 (Information Condition: Stereotypical, Counter-stereotypical) ANOVA. There was a main effect of prime,  $F(3, 1167) = 10.91, p < .001, \eta_p^2 = .027$ , which was not qualified by an interaction with information condition,  $F(3, 1167) = 1.02, p = .382, \eta_p^2 = .003$ . Participants skipped a higher proportion of trials that included Jonathan ( $M = .051, SD =$

.137) or Elizabeth ( $M = .054$ ,  $SD = .153$ ) primes than trials with control men ( $M = .034$ ,  $SD = .104$ ) or control women ( $M = .034$ ,  $SD = .105$ ) as primes. All differences in proportions between Jonathan and control trials, and Elizabeth and control trials, were significant (all  $ps < .01$ ), but the proportions skipped did not significantly differ between Jonathan and Elizabeth,  $t(387) = .854$ ,  $p = .394$ , Hedges'  $g_{av} = .02$ .

**Bimodality.** As was the case in Study 13, once more, the availability of an option to skip any trials in which the participant felt that their response to the pictograph might be affected by the prime did not eliminate bimodality in the distributions (Figure 35). For the proportion of pictographs judged as more related to doctors (vs. nurses) than average on Elizabeth and Jonathan trials in particular, the distributions appeared multimodal in both information groups.

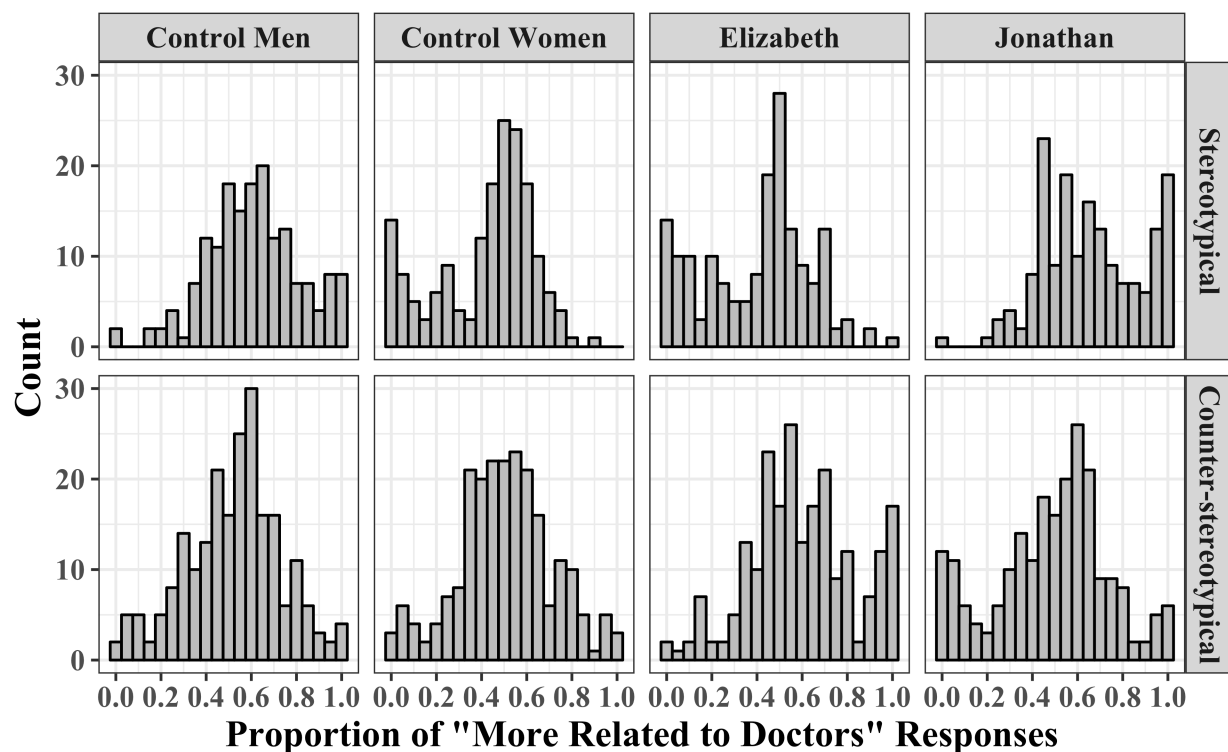


Figure 35. Frequency distributions of the proportion of pictographs judged to be more related to doctors (vs. nurses) than average in Study 14, by prime type and information condition.

Proportions were computed out of the total number of non-skipped trials.



## Discussion

The implicit impressions of the professions of Jonathan and Elizabeth on the PMP replicated the results of Study 12, showing that responses to those two individuals corresponded to the individuating information about their professions. These results, then, constitute additional evidence for a divergence between implicit impressions assessed via a misattribution measure and the IAT. Furthermore, the responses to the male and female control faces were also similar to Study 12, with implicit gender stereotyping in the stereotypical information condition but not in the counter-stereotypical information condition.

The findings of Study 14 also replicate the basic observation of bimodality in Study 13 in a different paradigm and using a different AMP-like task. Although participants were provided with an option to skip the trial whenever they felt that their response to the pictograph might be influenced by the prime image that came before it, and were encouraged to use that option, pictograph judgments within prime types were bimodal, particularly for Jonathan and Elizabeth. Furthermore, skipping was once again infrequent, with only slightly more than 5% of trials with Jonathan or Elizabeth as the prime being skipped. The frequency distributions showed that extreme responses to Jonathan and Elizabeth tended to be consistent with the individuating information that had been presented about each individual.

Together, Studies 13 and 14 show that in two different experimental paradigms, one in which participants learn about a seemingly negative person who becomes heroic and another in which participants learn stereotypical or counter-stereotypical professions for two individuals, bimodal frequency distributions emerge, despite the presence of an easy way to avoid any consciously-available influence of the primes on the targets – whether intentional or otherwise. Furthermore, this persistence was found on two different versions of the AMP: The standard

version that measures evaluative impressions (Study 13) and a modified version that measures impressions of professions (Study 14). The findings thus span both evaluative and semantic applications of the misattribution procedure (Gawronski & Ye, 2014), suggesting that bimodality might be more broadly common with such tasks. Importantly, the persistence of the pattern despite the skip instructions and the low frequency of skipping in general (see also Payne et al., 2013) best fits an interpretation of the bimodality as emerging under truly implicit misattribution, rather than a subset of participants eschewing the task instructions and intentionally rating the primes.

### **Study 15: Attention to an Irrelevant Dimension**

The preceding two experiments, coupled with the re-analyses of earlier data, begin to build the case that the bimodality in the distributions of the AMP data are not the result of participants intentionally rating the primes instead of the pictographs. Study 15 takes a different approach to establishing the implicitness of the AMP results by giving participants an attention-demanding distractor task to complete during the AMP that will divert attention away from the identities of the prime stimuli.

In light of the original concerns about the implicitness of the AMP raised by Bar-Anan and Nosek (2012), Gawronski and Ye (2015) tested whether the correlation between the AMP's priming effect and self-reported intentional responding – whether due to post-hoc confabulation or accurate reporting – could be reduced or eliminated by giving participants a distracting task to complete during the AMP that would call their attention away from the relevant dimension of the primes. This could helpfully reduce both true intentional responding by making it more likely that participants would respond intentionally based on the *distractor* task, and post-hoc confabulation of intentional responding to the critical prime dimension by making it more likely

that post-hoc inferences would instead be drawn about the distractor task that had occupied attention. Gawronski and Ye (2015) found that when participants were asked to tally the primes based on their races, a priming effect of age persisted but was no longer correlated with post-hoc intentional responding reports; likewise, when participants were asked to tally the primes based on their ages, a priming effect of race persisted but was also no longer correlated with post-hoc intentional responding reports. Self-reported intentional responding was instead only related to the dimension of the prime that was the focus of attention during the tally task.

Using a procedure adapted from the ideas presented in Gawronski and Ye (2015), Study 15 modified the AMP during a Francis West study to provide a tallying task designed to distract participants from the identity of the primes (Francis vs. control faces). This allowed a test of whether the revision effect would persist on the AMP under conditions of distraction, informing both the spontaneous (implicit) nature of the effect and providing more information on its efficiency under cognitive load (see Study 3). Additionally, the relationship between self-reported intentional responding and the Francis West priming effect, vs. priming based on the nature of the distractor task, can be examined.

A final goal of Study 15 is to examine whether the bimodal frequency distributions persist under conditions of considerable distraction, with no strong predictions as to whether it would. An elimination of the bimodality could occur if genuine intentional responding is curtailed, or if the load imposed by the distractor task reduces the priming effect; the latter would be consistent with work showing that processes often thought to be relatively automatic can be impaired by cognitive load (e.g., Gilbert & Hixon, 1991; Wells, Skowronski, Crawford, Scherer, & Carlston, 2011).

## **Method**

**Participants.** I aimed to recruit 300 participants from Mechanical Turk to participate in the study, receiving 301 responses. A priori exclusions consisted of: 1 server error, 9 participants who knew Mandarin and/or Cantonese, 2 participants who used a single key on every trial of the AMP, 7 for failing a manipulation check, and 2 for indicating that they had previously participated in a study using the Francis West story or a highly similar one. This left a final sample of 280 participants for analysis (52.9% women; Age  $M = 36.40$  years,  $SD = 11.49$ ).

**Procedure.** The experiment reproduced the basic Francis West paradigm (Study 1), with participants asked to read the standard story about Francis West that conveyed him as performing seemingly negative actions by breaking into his neighbors' homes. All participants then viewed the fire rescue information at Time 2, revealing that Francis had broken into the homes to save two children from a fire.

**The modified AMP.** The main differences between this study and the Francis West design from Chapter I dealt with the nature of the AMP, which was only administered once, at Time 2 (after the fire rescue information). First, to attempt to make it less likely that participants would notice frequent appearances of Francis West during the task, which might lead them to construe the task as being about Francis, the number of trials on the AMP was increased from 40 to 92, so that the number of Francis trials could remain similar to prior studies while allowing for Francis trials to be relatively infrequent. Francis appeared on 16 of the trials, with control faces on the remaining 76.

Second, to implement a procedure similar to Gawronski and Ye (2015), it was necessary to introduce a new dimension to each prime stimulus orthogonal to the identity as Francis vs. a control face, as the tallying dimensions in prior work were orthogonal (age vs. race in sets of young white, elderly white, young black, and elderly black faces) and any correlation between

the tallying dimension and the identity of the primes as Francis vs. a control face would risk keeping that prime dimension in the attention of participants, and also make it possible for reports of intentional responding to the dimension being tallied to introduce an artifactual correlation between those response reports and the critical prime effect.

To address these issues, the AMP trials were modified such that each prime face image was partially occluded by a black-and-white icon of a smile or frown, appearing to the left or right of the face shown in the prime image (but not blocking the face). Within the sets of Francis West and control face trials, 25% featured a smile icon on the left, 25% had a smile icon on the right, 25% had a frown icon on the left, and 25% had a frown icon on the right. These icons appeared at the same time as the prime image and left the screen with the prime image. The duration of the prime + icon was lengthened from the default 75ms to 200ms to give participants more time to view the icon. The tallying task required participants to pay attention to these icons while carrying out the main task of judging the pleasantness of the pictograph. There were two versions of the tallying task, with participants randomly assigned to complete one or the other. In one version, participants were asked to keep track of the number of icons appearing on each side of the prime over the course of the task; in the other version, participants were asked to keep track of the number of icons with each expression (smile vs. frown) appearing over the course of the task. They were directly told to ignore the other dimension of the icons.

The nature of the tally task was manipulated because it was unclear whether one would be more or less effective than the other. The dimensions featured in the tally task of Gawronski and Ye (2015) were social in nature (age and race), and it was possible that a relatively social tallying task (judging face-like expressions) would be a more effective distraction from the prime faces and/or could be necessary for preserving a strong main prime effect (Francis vs. control

faces) by keeping evaluation as a central focus of attention (Klauer & Musch, 2002; Spruyt, De Houwer, Hermans, & Eelen, 2007; cf., e.g., Ferguson, Bargh, & Nayak, 2005). On the other hand, because the smile-vs.-frown tallying task focused participants on a highly evaluative dimension besides the pictograph, there was some risk that it could interrupt misattribution of affect to the pictograph, so a left-vs.-right tallying task was also used as a precaution. The left vs. right tally task also supplemented another reason for introducing the variation in the side on which the icons appeared, which was to encourage participants to attend to the center of the prime image (the approximate location of the Francis/control face) in general given that they could not predict the side on which the icon would appear on any given trial.

After completing the modified AMP, participants were asked to report the results of the tally. These were not analyzed given the lack of any a priori predictions regarding the implications of tallying accuracy or the size of errors on this task, but the question was included to validate the cover story.

**Explicit evaluations.** Participants answered the same 6-item explicit liking scale employed in Study 1. Because of the central focus on implicit evaluations, analysis of the explicit scale is omitted in the present discussion.

**Reports of intentional and unintentional responding.** Next, participants were asked two questions, one to obtain their belief about the extent to which they had intentionally judged the primes on the AMP, and the second to obtain their belief about the extent which their responses to the pictographs may have been unintentionally influenced by the primes (both adapted from Payne et al., 2013). To gauge intentional responding, participants were asked, “During the task in which you were asked to rate the pleasantness of the Chinese pictographs (symbols), did you intentionally rate the images of people or face icons instead of the symbols?”

and responded on the following 5-point scale: *Not at all, I rated the symbols; Usually no; Sometimes, but not always; Usually yes; Yes, I rated the images of people or face icons.* To assess reports of unintentional influence, participants were asked, “During the task in which you were asked to rate the pleasantness of the Chinese pictographs (symbols), were your ratings of the symbols unintentionally influenced by the images of people or face icons?” and responded on the following 5-point scale: *Not at all; Usually no; Sometimes, but not always; Usually yes; Yes.*

On in the intentional responding question, if participants selected anything other than “Not at all, I rated the symbols”, they were presented with an intermediate page between these two questions that followed up on their report of intentional responding. On this page, they were asked to *Agree* or *Disagree* (dichotomous choice) with 7 potential reasons for responding intentionally, which had been drafted by the research team. They could agree with as few or as many of the 7 reasons as they liked:

- 1) *I found it more interesting to rate the images of people or face icons.*
- 2) *I thought I was supposed to rate the images of people or face icons, not the Chinese pictographs.*
- 3) *I tried to rate the Chinese pictographs, but kept rating the images that flashed beforehand by accident.*
- 4) *I wanted to help the researchers find a connection between my responses and the images.*
- 5) *I had too much trouble seeing the Chinese pictographs, so I rated the other images instead.*
- 6) *I couldn't pay enough attention to the Chinese pictographs while also trying to tally the face icons.*
- 7) *It was easy to just press the key that was on the same side as the face icons.*

Secondly, participants were given an open-ended response box in which they could offer any additional reasons for which they might have intentionally judged the images of faces or icons instead of the pictographs, if not captured by the reasons available above. These open-ended responses have not yet been systematically analyzed and coded, and so will not be discussed further here.

**Final questions.** Participants were additionally asked two exploratory questions, one on how difficult they found the tallying task to be from 1 (*Not at all difficult*) to 7 (*Extremely difficult*) and one asking the extent to which they tried to keep an accurate tally of the face icons, from 1 (*Did not try at all*) to 7 (*Tried as hard as I could*). They then answered the story comprehension questions from the studies in Chapter II, were asked to identify Francis West in photo array, identified the final information that they had learned about Francis from a set of 3 choices, indicated if they had participated in a study using the Francis West story before, and provided demographic information.

## Results

**Implicit evaluations.** The proportion of pictographs judged as more pleasant than average was computed for each participant within each cell of a 2 (Prime Type: Francis West, control faces) x 2 (Icon Expression: frown, smile) x (Icon Side: left, right) x 2 (Tally Task: expression, side) design. These proportions were analyzed within a mixed ANOVA, with the last factor manipulated between-participants and the rest manipulated within-participants.

A main effect of prime type obtained,  $F(1,278) = 13.18, p < .001, \eta_p^2 = .045$ , such that Francis West was implicitly more positive ( $M = .58, SD = .21$ ) than control faces ( $M = .54, SD = .14$ ). This main effect was not qualified by interactions with tally instruction,  $F(1,278) = 1.82, p = .179, \eta_p^2 = .006$ , icon expression,  $F(1,278) = .88, p = .349, \eta_p^2 = .003$ , or icon side,  $F(1,278) =$



.80,  $p = .372$ ,  $\eta_p^2 = .003$ . There was a three-way interaction between prime, icon side, and icon expression,  $F(1,278) = 7.50$ ,  $p = .007$ ,  $\eta_p^2 = .026$ , to be discussed below. The four-way moderation of that effect by tally instruction was not significant,  $F(1,278) = .825$ ,  $p = .365$ ,  $\eta_p^2 = .003$ .

There were main effects of icon side,  $F(1,278) = 24.38$ ,  $p < .001$ ,  $\eta_p^2 = .081$ , and icon expression,  $F(1,278) = 61.99$ ,  $p < .001$ ,  $\eta_p^2 = .182$ , which were qualified by an interaction between the two,  $F(1,278) = 8.12$ ,  $p = .005$ ,  $\eta_p^2 = .028$ . Collectively, the interpretation of these effects is that pictographs were more likely to be judged as less pleasant when the face icon appeared on the left side (which was the location of the “less pleasant” response key) than the right side (the location of the “more pleasant” key), and that pictographs were more likely to be judged as more pleasant than average when the accompanying face icon was smiling rather than frowning. However, the effect of side was stronger for frowning icons than smiling icons, and the effect of icon expression was stronger when the icons appeared on the left side (the side of the “less pleasant” key).

As previewed above, however, these effects were subsumed by a significant three-way interaction between prime, icon side, and icon expression. This interaction was such that the prime effect (Francis vs. control faces) was strongest when the icon expression mismatched the side on which it appeared; in other words, Francis West was significantly more implicitly positive than control faces on trials in which a smiling icon appeared on the left side (with the “less pleasant” key),  $t(279) = 4.01$ ,  $p < .001$ , Hedges’  $g_{av} = .252$ , or on trials in which a frowning icon appeared on the right side (with the “more pleasant” key),  $t(279) = 3.35$ ,  $p = .001$ , Hedges’  $g_{av} = .211$ . The prime effect was smaller, though still significant, on trials in which smiling icons appeared on the right side,  $t(279) = 2.19$ ,  $p = .029$ , Hedges’  $g_{av} = .149$ , and did not reach

significance on trials in which frowning icons appeared on the left side,  $t(279) = .621$ ,  $p = .535$ , Hedges'  $g_{av} = .037$ . Figure 36 shows the AMP scores within each cell of the interaction between icon side, icon expression, and prime type.

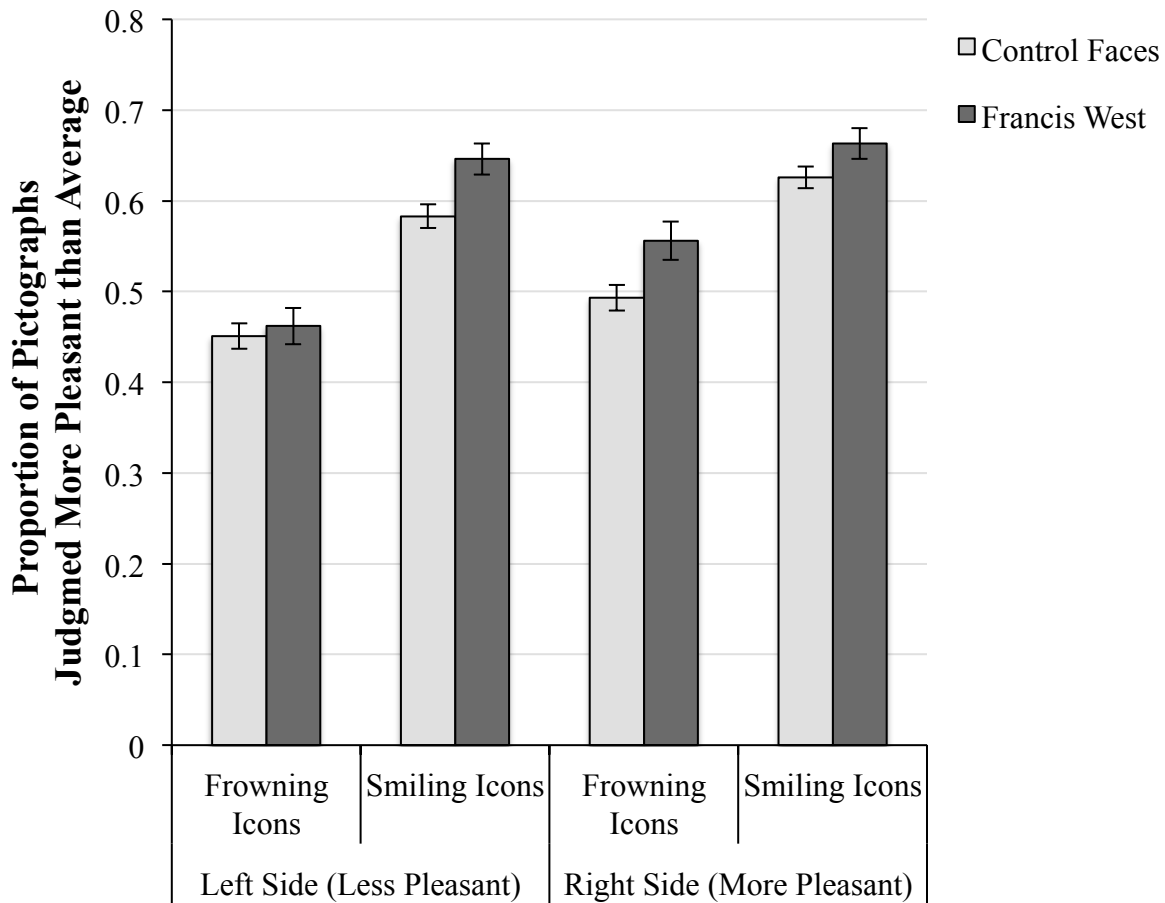


Figure 36. Proportion of pictographs judged more pleasant than average in Study 15, by icon side, icon expression, and prime type. Error bars are standard errors.

**Bimodality.** As shown in Figure 37, the frequency distribution of AMP proportions on Francis West trials was far closer to normal than in prior studies, with only a somewhat fatter tail on the upper end of the distribution. The modifications made to the study seem to have been effective in altering that otherwise prominent pattern.

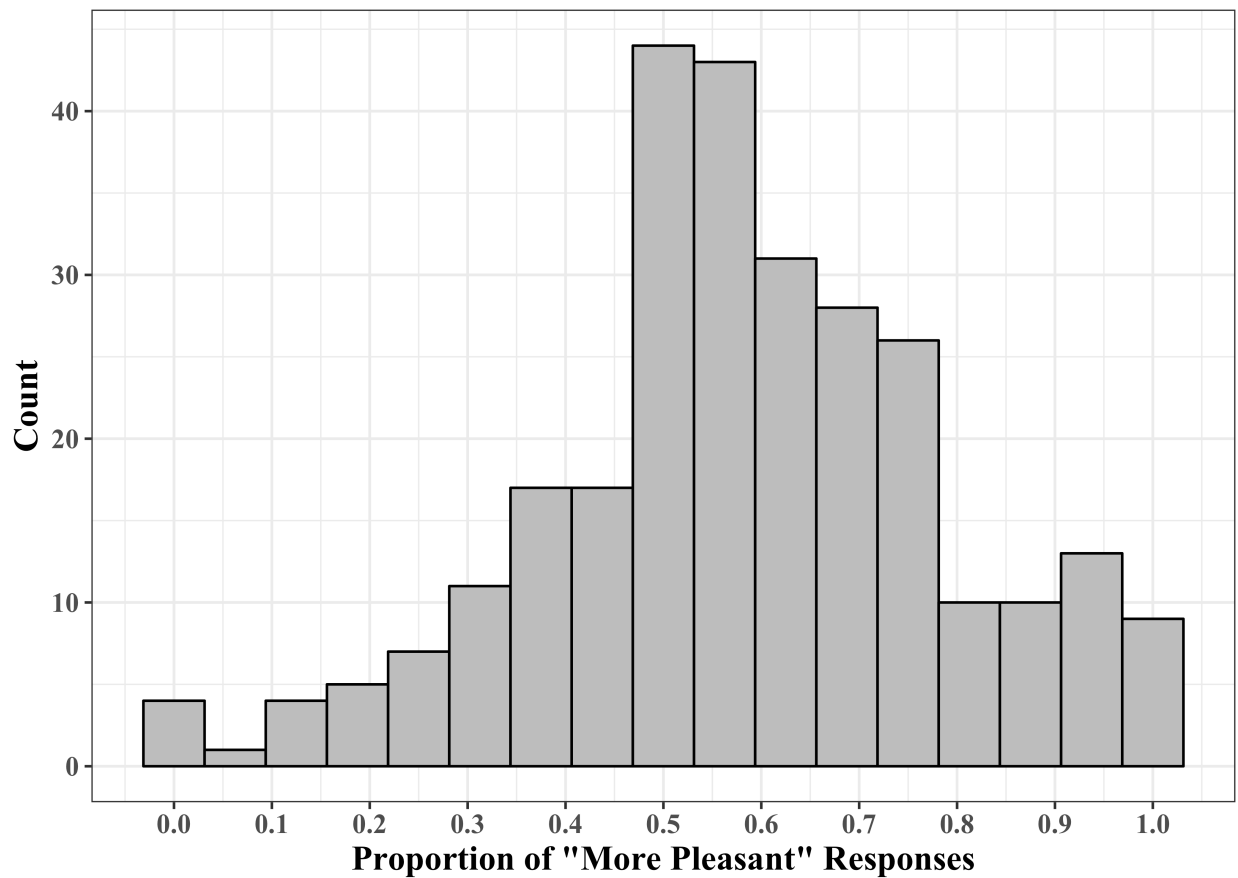


Figure 37. Frequency distribution of the proportion of pictographs judged to be more pleasant than average on trials in which Francis West was the prime image, in Study 15.

**Self-reported intentional and unintentional influence.** I next turned to an analysis of the two questions asking participants to report, first, the degree to which their responses to the pictographs were *intentionally* affected by the prime faces that came before each, and second, the degree to which their responses had been *unintentionally* affected by those primes. Mean unintentional responding ( $M = 2.54$ ,  $SD = 1.20$ ) was higher than mean intentional responding ( $M = 2.08$ ,  $SD = 1.27$ ),  $t(279) = 7.02$ ,  $p < .001$ , Hedges'  $g_{av} = .370$ . The two measures were strongly positively correlated,  $r(278) = .609$ ,  $p < .001$ , which is consistent with the proposal that

intentionality reports after completing the AMP may represent post-hoc inferences, given that the same participants who report responding intentionality also report unintentional influence; awareness of some connection between the primes and their responses may prompt participants to make either inference when a plausible hypothesis to explain the connection is provided in the text of the question (Payne et al., 2013).

The intentionality and unintentionality reports were next examined for evidence of correlation with three difference-score measures tapping different priming effects within the AMP: the relative implicit positivity of Francis West vs. control faces, the relative implicit positivity of smiling icons vs. frowning icons, and the relative implicit positivity of icons appearing on the right side vs. icons appearing on the left side. Intentionality reports were correlated only with the implicit preference for smiling icons over frowning icons,  $r(278) = .325$ ,  $p < .001$ , and not with implicit positivity of Francis West over control faces,  $r(278) = .052$ ,  $p = .387$ , or with implicit evaluations of right-side icons over left-side icons,  $r(278) = .041$ ,  $p = .496$ . Reported *unintentional* influence was correlated with implicit preference for smiling icons over frowning icons,  $r(278) = .238$ ,  $p < .001$ , as well as implicit evaluations of Francis West vs. control faces,  $r(278) = .170$ ,  $p = .004$ . These reports were not correlated with implicit right-side vs. left-side icon preference,  $r(278) = .081$ ,  $p = .177$ .

I next ran a multiple linear regression analysis on implicit evaluations of Francis West vs. control faces, including both reported intentional and unintentional responding as predictors. This analysis showed a unique effect only of unintentional response reports,  $\beta = .220$ ,  $t(277) = 2.96$ ,  $p = .003$ , and not intentional response reports,  $\beta = -.082$ ,  $t(277) = 1.11$ ,  $p = .269$ . A separate regression using the same predictors but implicit preference for smiling vs. frowning icons found the opposite pattern, with a unique effect only of intentional response reports,  $\beta =$

.287,  $t(277) = 4.01$ ,  $p < .001$ , and not unintentional response reports,  $\beta = .063$ ,  $t(277) = .88$ ,  $p = .380$ .

**Reported reasons for intentional responding.** If participants selected anything other than the lowest scale endpoint – “Not at all, I rated the symbols” – on the question asking whether they ever intentionally rated the faces during the AMP, they were then presented with seven possible reasons for such intentional responding and asked to select “agree” or “disagree” to indicate whether each reason contributed to their behavior on the task. The levels of agreement with each reason (in descending order) among participants who indicated some intentional responding (53.21% of the total sample) were: Inability to pay enough attention to the pictographs while also tallying the face icons (63.1%), wanting to help the researchers find a connection between the responses and the images (57.7%), trying to rate the pictographs but judging the images that came before by accident (57.0%), finding the images of people or face icons more interesting to rate (36.2%), having too much trouble seeing the pictographs (35.6%), finding it easier to press the key on the same side as the face icon (21.5%), and thinking that the images of people or face icons were supposed to be rated (17.4%). A series of point-biserial correlations testing for relationships between endorsement of each of these reasons and implicit preference for Francis West vs. control faces found that only one – reported reliance on the icon side in informing responses – was related to the implicit preference, but in the negative direction,  $r(147) = -.199$ ,  $p = .015$ . That is, to the degree that participants reported relying on the side of the screen on which the smiling/frowning icon appeared in determining their judgments on each trial of the AMP, the less implicitly positive Francis West was relative to control faces. Implicit preference for smiling vs. frowning icons, on the other hand, was related to agreement with more of the reasons, including a desire to help the researchers,  $r(147) = .192$ ,  $p = .019$ , difficulty

seeing the pictographs clearly,  $r(147) = .187, p = .022$ , and reliance on icon side,  $r(147) = .161, p = .05$ . Finally, implicit preference for right-side vs. left-side icons was (unsurprisingly) related only to reported reliance on icon side,  $r(147) = .236, p = .004$ .

## **Discussion**

The most central finding in Study 15 is that the revision effect of Francis West becoming more implicitly positive than control faces after participants learn the fire rescue revelation replicated under conditions of heavy distraction. Though a smaller effect than in the studies contained within Chapter II (which is perhaps not surprising given the taxing nature of attending to two running tallies over the course of the AMP), implicit preference for Francis West survived an elimination of bimodality in the distribution of AMP scores and was not correlated with subjective reports of intentional responding or endorsement of all but one (in a non-troubling direction) of the reported reasons for responding intentionally. Instead, paralleling the results of Gawronski and Ye (2015), reports of intentional responding correlated with priming effects along the dimension to which the participants attended: the properties of the icons.

Interestingly, reported extent of unintentional influence correlated with implicit evaluations of Francis West over control faces, even though reported intentional responding did not. Though not predicted, this relationship could suggest that participants have some more fine-grained insight into their own implicit cognition (see also Hahn et al., 2014), or could be produced by a lay inference that their judgments were probably influenced by the appearance of Francis to the degree that they feel strongly about him, even without conscious memory for any actual such influence.

Among the slightly more than half of participants who reported some amount of intentional responding, endorsement of many of the potential reasons for doing so was quite high

(3 were endorsed by more than 50% of participants indicating some amount of intentional responding). It is difficult to know from the present study the degree to which endorsement of any of these reasons might have been increased by the presence of the tallying task, which would limit generalization from these results to earlier work; reasons like finding it difficult to pay attention to the pictographs and accidentally judging the earlier stimuli almost certainly were increased by the presence of the tallying task (and the latter reason could even be considered *unintentional* responding). More generally, as with providing participants with the very idea of intentional responding and asking for them to reflect on whether that occurred, it is possible that providing these reasons produced some confabulation (Payne et al., 2013).

Whereas Studies 13-14 examined whether bimodality may be result of intentional responding, Study 15 found evidence that the attenuation of the bimodality through a distracting and taxing task does not keep new information that provides a reinterpretation of Francis West's earlier actions from producing a significantly positive implicit impression of him.

Due to the difficulty of interpreting the self-report measures of intentional and unintentional responding in this study, and the cognitive demands imposed by the tally task that may both reduce real implicit priming effects and make it less practical to put participants through multiple iterations of this version of the AMP, the next two studies take a different approach. First, Studies 16-17 focus squarely on reducing bimodality (whether it is or is not due to intentional responding) rather than de-correlating the AMP priming effect from intentional response reports, given that the former has been the original impetus for concern and the latter is known to be an inherently ambiguous and probably misleading indicator of actual intentional responding (Gawronski & Ye, 2015; Payne et al., 2013). Second, these next studies will attempt to achieve the reduction in bimodality by increasing the vigilance of participants against any

temptation to intentionally judge the pictographs on the basis of the primes, rather than just giving them a way to skip the trial if the influence would occur and encouraging them to use it (Studies 13-14). Third, the studies will use the paintings stimuli set made use of in the hunting podcast study (Study 10), which found weaker implicit impression change in a real-world context but distributions closer to normal. The paintings were used here because it is possible that although they are all similar in many ways, with their varied colors and patterns the paintings might carry greater inherent evaluative preferences for the participants, such that it would be less likely for a strong prime to “swamp” the responses to the paintings in participants who are attempting to perform the task properly. Additionally, it was thought possible that genuine intentional responders might find the paintings more interesting or easier to rate than pictographs, and easier to see (with their varied, bright colors), which might coax such participants into following the task instructions to judge the paintings instead of the primes.

Given the importance of the comparison between the fire rescue (reinterpretation) and subway rescue (non-reinterpreting positive) information conditions presented in Study 2 for establishing that the propositional relationship between the new and old information matters greatly in determining how much revision in an initial negative implicit evaluation the new information will produce, Studies 16 and 17 will both include the three information conditions used in Study 2: control, fire rescue, and subway rescue.

### **Study 16: Warning to Avoid Influence – I**

#### **Method**

**Participants.** I aimed to recruit 750 participants from Prolific Academic, to achieve approximately 250 per each between-participants condition. Out of 750, I made the following a priori exclusions: 1 for failing to complete all parts of the experiment, 3 for server errors, 8 for



failing a manipulation check, 4 for misidentifying Francis West in the photo lineup, 59 for pressing a single key on every trial of at least one AMP, and 27 for indicating prior participation in a study using the Francis West story (or a highly similar one). The last criterion was necessary to weed out participants who had participated in a related study on the Mturk platform. This resulted in a final sample size of 648 (53.7% women, Age  $M = 33.01$  years,  $SD = 11.44$ ).

**Procedure.** Participants completed the standard Francis West procedure including the information conditions from Study 2 (control, fire rescue, and subway rescue). They completed the AMP twice, once after reading the initial information at Time 1 (when Francis was ostensibly negative) and once after reading the final screen of information about him (Time 2). For all participants, the AMPs included the original configuration of 40 trials (20 with Francis West as the prime and 20 with control faces as the prime), but were modified in two ways. First, the painting stimuli from Study 10 were used, in two sets of 40, with painting set order manipulated between-participants. Second, all participants viewed an additional, final page of instructions on the AMP, which implored them to avoid intentionally judging the primes instead of the pictographs. To avoid scaring participants into going to lengths to avoid *any* connection between the primes and the pictographs, including monitoring for any unintentional prime effects or ensuring 50% pleasantness responses to each prime type, the page also told them that it was fine if they noticed that their responses occasionally matched how they *would have* rated the faces, but that they should just focus on rating the paintings. The appearance of this final page of information is shown in Figure 38.

# IMPORTANT!

It is **CRUCIAL** that you try to evaluate only the paintings, rather than the faces that come before them.

Even if you are trying to rate only the paintings, you may notice occasionally that your ratings of the paintings are consistent with how you *would have rated* the previous faces. **That is totally okay.** You do not have to worry about whether your ratings of the paintings are consistent, or not, with how you feel about the faces.

Your only job is to evaluate the paintings, and we really need your help in doing this. Do not even worry about the faces, or how your painting judgments seem to occasionally match the faces.

Your help to this research is **really** important. Our data depend on participants like you trying as hard as possible to follow instructions. We really need and appreciate your help.

***Thank you so much for your help!***

The task will now begin. When you are ready, place your fingers on the 'd' (less pleasant) and 'k' (more pleasant) keys, and press the space bar.

***d key = less pleasant***

***k key = more pleasant***

Figure 38. The final page of instructions on the AMP in Studies 16-17.

Explicit evaluations were assessed after each AMP, with the 6-item scale used throughout this work. At the end of the experiment, participants completed a manipulation check asking them to identify Francis' final action out of 6 choices, identified him in a photo lineup, reported if they had previously participated in any experiment using the same Francis West paradigm or a highly similar one, and provided demographic information.

## Results

**Implicit evaluations.** The proportion of paintings judged to be more pleasant than average was computed at each time within each prime type, and analyzed in a 2 (Time: Time 1, Time 2) x 2 (Prime Type: Francis West, Control Faces) x 3 (Information Condition: Control, Fire Rescue, Subway Rescue) mixed ANOVA, with the first two factors manipulated within-subjects.

The three-way interaction between time, prime, and final information was significant,  $F(2,645) = 6.84, p = .001, \eta_p^2 = .021$ . Figure 39 shows mean implicit evaluations within each cell of this design.

**Time 1 initial formation.** I first examined whether Francis West was implicitly negative relative to control faces after the initial story portraying him as breaking into and damaging the homes of his neighbors. In the overall sample, there was a marginally significant effect of prime at Time 1,  $t(647) = 1.72, p = .085$ , Hedges'  $g_{av} = .069$ . A between-participants ANOVA found no evidence that the formation (prime) effect varied by subsequent information condition, which was not yet manipulated,  $F(2,645) = 1.48, p = .228, \eta_p^2 = .005$ .

**Time 2 revision.** After participants learned the final information about Francis West (Time 2), there was a significant interaction between prime and information condition,  $F(2,645) = 16.16, p < .001, \eta_p^2 = .048$ . This interaction was decomposed for follow-up tests into two orthogonal interaction contrasts, one comparing the prime effect in the control condition to the prime effect in the other two combined (fire rescue and subway rescue), and the other comparing the prime effect in the fire rescue condition to that in the subway rescue condition. The first contrast was statistically significant,  $F(1,645) = 28.42, p < .001, \eta_p^2 = .042$ , as was the second,  $F(1,645) = 4.10, p = .043, \eta_p^2 = .006$ .

Simple-effects tests on the prime effects at Time 2 indicated that in the control condition, Francis West remained significantly less implicitly positive ( $M = .44, SD = .20$ ) than control

faces ( $M = .50$ ,  $SD = .21$ ),  $t(210) = 3.70$ ,  $p < .001$ , Hedges'  $g_{av} = .263$ . In the fire rescue condition, however, Francis West had become implicitly more positive ( $M = .50$ ,  $SD = .21$ ) relative to the control faces ( $M = .45$ ,  $SD = .20$ ),  $t(214) = 4.19$ ,  $p < .001$ , Hedges'  $g_{av} = .268$ . Finally, in the subway rescue condition, implicit evaluations of Francis West ( $M = .46$ ,  $SD = .20$ ) and control faces ( $M = .45$ ,  $SD = .20$ ) did not differ,  $t(221) = 1.14$ ,  $p = .254$ , Hedges'  $g_{av} = .076$ .

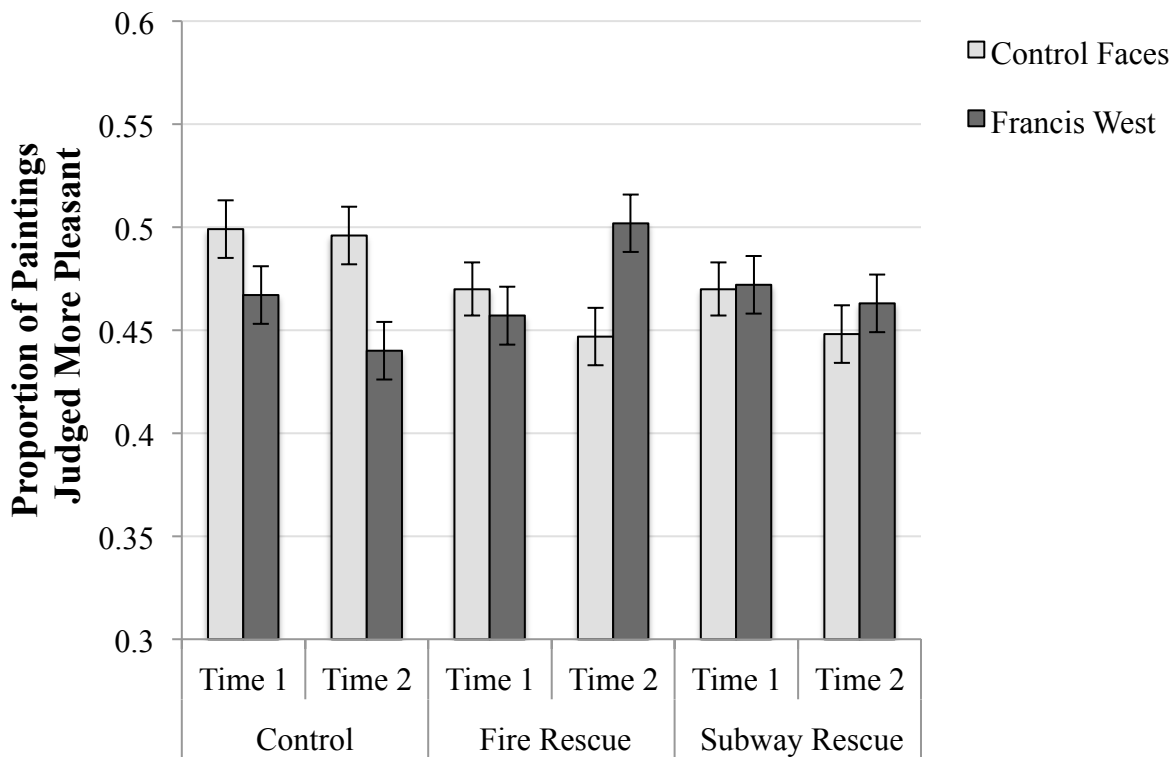


Figure 39. Mean proportion of paintings judged more pleasant than average in Study 16, by information condition, time, and prime type. Error bars are standard errors.

**Bimodality.** Figure 40 shows the frequency distribution at Time 1 for trials in which Francis West was the prime stimulus, and Figure 41 shows the frequency distributions at Time 2

within each information condition for trials in which Francis West was the prime stimulus, with no bimodality present at Time 1 or at Time 2 in any of the three conditions.

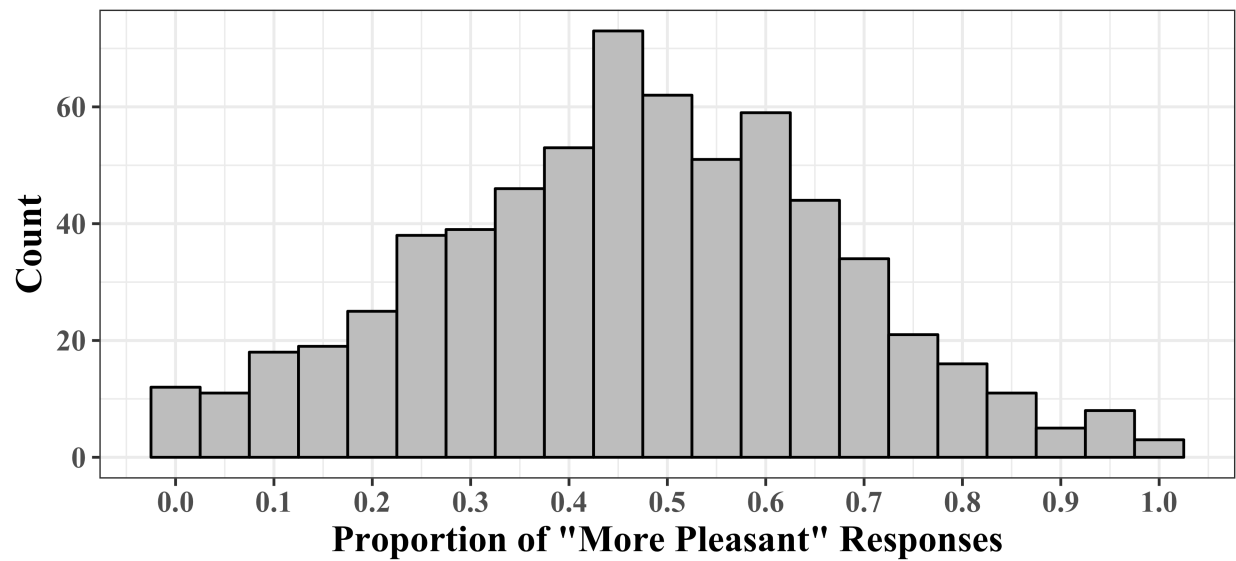
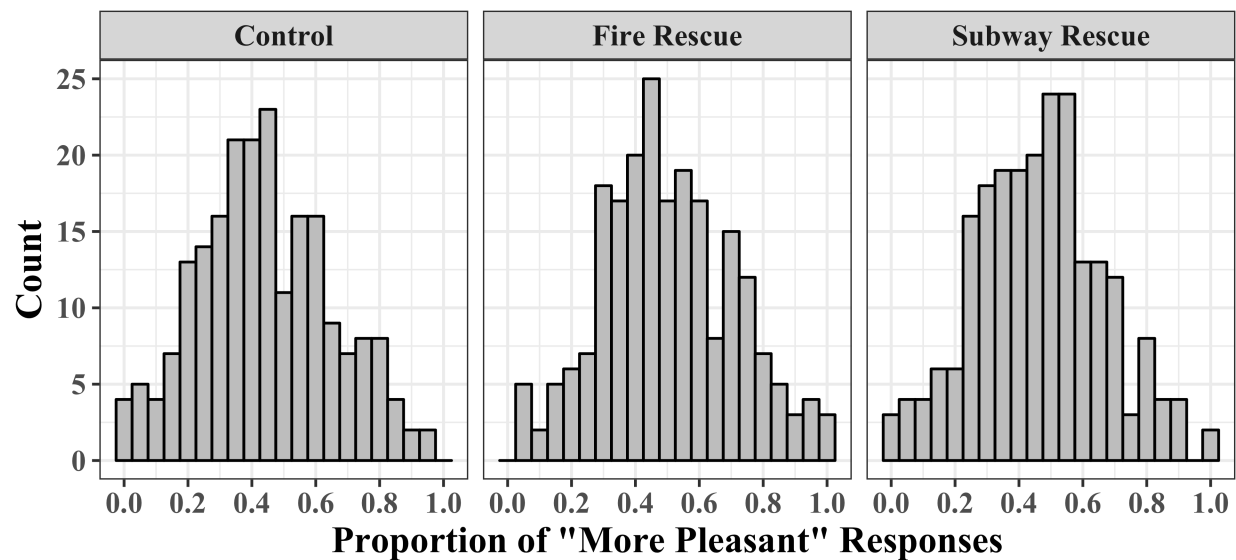


Figure 40. Frequency distribution of the proportion of paintings judged to be more pleasant than average on trials in which Francis West was the prime image at Time 1 in Study 16.



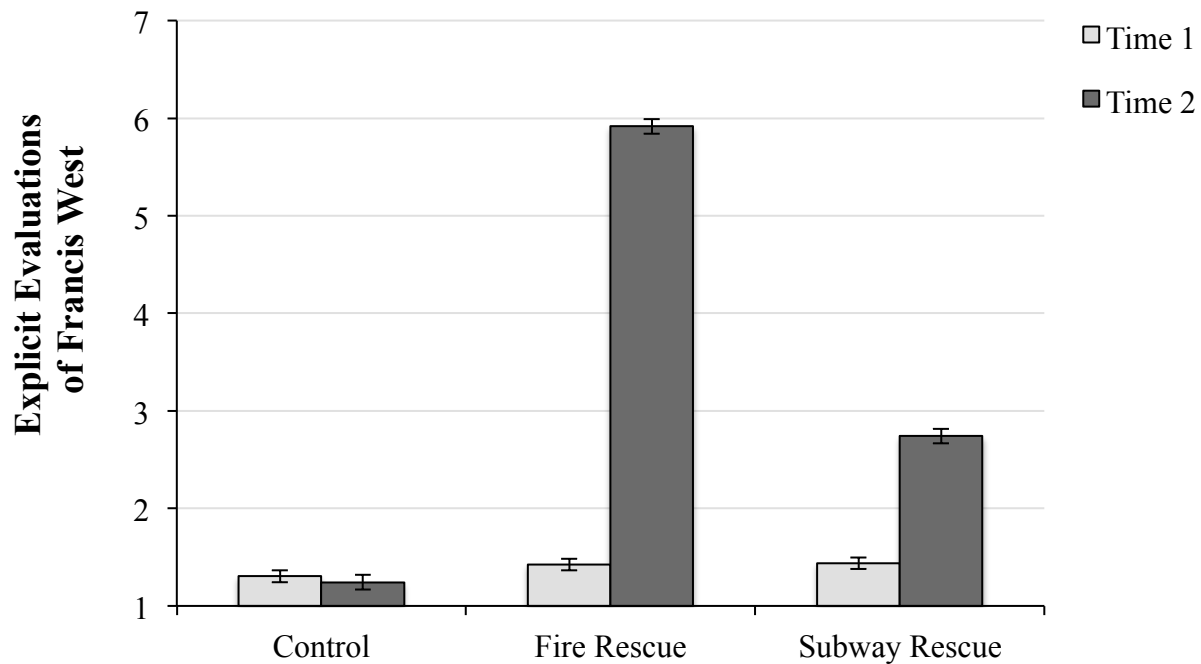
*Figure 41.* Frequency distributions of the proportion of paintings judged to be more pleasant than average on trials in which Francis West was the prime image in Study 16, by information condition.

**Ancillary analysis.** Given the relatively weak initial formation of negative implicit evaluations of Francis West in this study, I conducted an ancillary analysis (similar to Study 1b) to examine the effects of information condition at Time 2 exclusively among those participants who showed negative implicit evaluations of Francis West at Time 1 (i.e., people for whom the proportion of paintings judged pleasant on trials in which Francis West was the prime image was lower than the proportion of paintings judged pleasant on trials in which a control face was the prime image). Though no measure is process-pure (Conrey et al., 2005; Payne et al., 2001, Payne et al., 2010; Sherman et al., 2008), the idea that this procedure will select for those participants with genuine initial negative evaluations of Francis West is supported by two indications of genuine signal (vs. solely noise) in the measure: (a) the correlation between explicit liking and implicit preference for Francis West (the difference score of proportion of “pleasant” painting responses on Francis trials minus the proportion on control trials) at Time 1 was positive and significant in the overall sample,  $r(646) = .177, p < .001$ , and within the fire rescue condition,  $r(213) = .163, p = .016$ , and subway rescue condition,  $r(220) = .225, p = .001$ ; the correlation was marginal in the control condition,  $r(209) = .120, p = .083$ ; and (b) similar to a finding in the initial line of studies (Chapter II), the correlation between Time 1 and Time 2 implicit preference for Francis West over control faces was significant in the control condition,  $r(209) = .385, p < .001$ , and subway rescue condition,  $r(220) = .235, p < .001$ , but not the fire rescue condition,  $r(213) = -.007, p = .923$ , which is theoretically consistent with the idea that

only in the fire condition will reinterpretation completely sever the relationship between initial and final implicit evaluations of Francis.

The ancillary analysis – with this restricted sample – corroborated the results of the main set of tests. There remained a significant interaction between prime and information condition at Time 2,  $F(2,300) = 13.18, p < .001, \eta_p^2 = .081$ , and both the control vs. other contrast and fire rescue vs. subway rescue contrasts were significant,  $F(1,300) = 18.20, p < .001, \eta_p^2 = .057$ , and  $F(1,300) = 8.23, p = .004, \eta_p^2 = .027$ , respectively. In the control condition, Francis West was less implicitly positive ( $M = .42, SD = .22$ ) than control faces ( $M = .51, SD = .22$ ),  $t(103) = 3.64, p < .001$ , Hedges'  $g_{av} = .407$ . In the fire rescue condition, Francis West had become more implicitly positive ( $M = .50, SD = .22$ ) than control faces ( $M = .43, SD = .21$ ),  $t(98) = 3.26, p = .002$ , Hedges'  $g_{av} = .315$ . Lastly, in the subway rescue condition, Francis West ( $M = .45, SD = .19$ ) and the control faces ( $M = .47, SD = .19$ ) did not significantly differ,  $t(99) = 1.13, p = .260$ , Hedges'  $g_{av} = .114$ .

**Explicit evaluations.** Analyzing the explicit evaluation mean scores at both Time 1 and Time 2 across information conditions revealed a significant interaction between time and information condition,  $F(2,645) = 949.54, p < .001, \eta_p^2 = .746$ , with results that parallel those of Study 2. Figure 42 displays the mean explicit evaluations of Francis West in each condition.



*Figure 42.* Mean explicit evaluations of Francis West in Study 16, by time and information condition. Error bars are standard errors.

## Discussion

The results of Study 16 found that the critical difference between the fire rescue and subway rescue information conditions, first established in Study 2, replicated under conditions that eliminated the bimodal distributions in the AMP data. The argument that reinterpretation is a mechanism through which the revision of implicit evaluations can occur was bolstered by studies in which self-reported use of reinterpretation mediated change (Studies 4, 5, and 10), but the experimental difference between the fire rescue and subway rescue information conditions is arguably the most straightforward and clearest evidence for the important role played by the connection between the new information and the old. Though both actions were rated as equivalently positive out of the context of the rest of the story, only the heroic act that recasts the earlier story in a negative light produced a reversal in implicit evaluations in those initial studies.



The replication of the difference in revision between the fire rescue and subway rescue conditions is thus a critical step in demonstrating that regardless of the ultimate meaning of the bimodal pattern in the initial AMP data, the role of reinterpretation in reversing implicit evaluations is not dependent on the presence of that distribution (which was also suggested by the re-analyses of earlier studies showing that the revision effect replicated even when excluding the uniform responders).

The outcome of this study demonstrates that the combination of adjustments to the AMP, including the additional page of warning instructions and the use of painting stimuli instead of pictographs, effectively eliminated bimodality without imposing a taxing and distracting task on participants (Study 15). It is unknown at present whether one of these elements would be sufficient to achieve the elimination of bimodality or whether both are required; nonetheless, the study offers a possible route through which future studies using the AMP can avoid bimodal distributions in the data, if desired.

One difference of potential interest between the present results and earlier studies that included the subway rescue condition is that here, final implicit evaluations toward Francis in the subway rescue condition were indistinguishable from control faces, whereas in earlier studies Francis West remained significantly *negative* relative to control faces after participants learned this control heroic information. Though the meaning of this difference is open to interpretation, it suggests that the negativity that remained toward Francis in this condition in earlier work may have been somewhat tenuous, not able to survive the changes made to the AMP in Study 16. Nonetheless, the most critical finding was that only in the fire rescue information condition were final implicit evaluations of Francis West significantly positive.

The results were not without limitations, however. The initial formation of negative implicit evaluations of Francis West was surprisingly weak, especially given that initial formation has typically been robust. Indeed, the elimination of bimodality seems to have reduced the size of the priming effects across the board, but this reduction was most pronounced at Time 1. The interpretation of this difference from earlier studies is ambiguous, because although it would be expected if “true” intentional responders (with their inflated effect sizes) are now less frequent, it also could occur if participants who are faithfully attempting to follow directions and would normally have extreme scores were prompted by the warning instruction to engage in some monitoring of their responses to avoid strong correlations between the primes and their responses (despite our statement that unintentional correlation is perfectly acceptable). The use of the paintings as target stimuli could also have reduced priming effects to the extent that participants might have stronger evaluative preferences among them.

Fortunately, the ancillary analysis, paralleling the procedure employed in Study 1b with the similarly weaker initial formation effects found on the IAT, found that revision occurred in the fire rescue condition even among only those participants who *did* show initially negative implicit evaluations of Francis West at Time 1, and also found that the difference between the fire rescue and subway rescue conditions remained. Analyses including only participants who show initial formation of a response are arguably even more appropriate than use of the whole sample in work focused on examining *change*, as formation is a necessary antecedent to change and analyses of whole samples will include participants who failed to produce that initial response. Excluding participants who do not show initial formation is for this reason common in other work meant to examine change, such as studies of memory reconsolidation (e.g., Schiller, Kanen, LeDoux, Monfils, & Phelps, 2013; Steinfurth et al., 2014).

Even with the ancillary analysis, however, it would be ideal to demonstrate that stronger initial formation, detectable in the full sample, can still occur and be undone through reinterpretation with the modified AMP that eliminates bimodality. The aim of Study 17 is to replicate the results of Study 16 with stronger initial formation.

### **Study 17: Warning to Avoid Influence - II**

#### **Method**

**Participants.** As with Study 16, I aimed to recruit 250 participants per each between-participants condition (750 in total) from Prolific Academic. The following exclusions were determined a priori: 2 for not finishing the study, 1 for a server timeout, 32 for indicating participation in a prior related study, 10 for failing a manipulation check, 5 for misidentifying Francis West in a photo lineup, and 81 for using a single key on every trial of at least one AMP. This brought the final sample to 619 participants for analysis (59.3% women, Age  $M = 31.37$  years,  $SD = 10.82$ ).

**Procedure.** The participants completed a procedure identical to Study 16, with one modification: For all participants, an additional detail was added to the initial story about Francis West at the very beginning, noting that although Francis's town is 99% white, the Griffins and Wards had just moved in, and were the first interracial couples in town. This is the same additional detail that had been used in Study 1b, and was meant to lead participants to infer that the actions undertaken by Francis in breaking into the homes of those neighbors might have been motivated by racial hatred, thereby increasing initial negativity felt toward Francis. The only other alteration to the story (as in Study 1b) was to change "years worth of careful cultivation" (in reference to a garden maintained by Francis's neighbor) to "countless hours of careful cultivation" for consistency with the claim that the neighbors had moved to town only recently.

As in Study 16, participants were assigned to one of the control, fire rescue, or subway rescue information conditions, and completed all of the same measures in an identical fashion to the protocol from that study.

## Results

**Implicit evaluations.** Proportions of pleasant responses were computed following the procedure for Study 16, and analyzed within a 2 (Time: Time 1, Time 2) x 2 (Prime Type: Francis West, Control Faces) x 3 (Information Condition: Control, Fire Rescue, Subway Rescue) mixed ANOVA, with the first two factors manipulated within-participants. As in Study 16, the three-way interaction between time, prime, and information condition was significant,  $F(2,616) = 6.59, p = .001, \eta_p^2 = .021$ .

**Time 1 initial formation.** At Time 1, unlike in the prior study, there was evidence of significant formation of negative implicit evaluations of Francis West relative to control faces,  $t(618) = 3.74, p < .001$ , Hedges'  $g_{av} = .160$ . An ANOVA showed that these formation prime effects did not significantly differ between (not-yet-assigned) information conditions,  $F(2,616) = .29, p = .750, \eta_p^2 = .001$ .

**Time 2 revision.** A significant interaction between prime and information condition emerged on the Time 2 AMP scores gauging implicit evaluations after participants learned the final information about Francis West (specific to their information condition),  $F(2,616) = 11.68, p < .001, \eta_p^2 = .037$ . As with Study 16, I decomposed the Time 2 effect into orthogonal contrasts testing for differences in the prime effect in the control condition vs. the other two, and between the fire rescue and subway rescue conditions. The first contrast was significant,  $F(1,616) = 17.71, p < .001, \eta_p^2 = .028$ , as was the second,  $F(1,616) = 5.13, p = .024, \eta_p^2 = .008$ . Figure 43 displays the mean implicit evaluations within each condition.

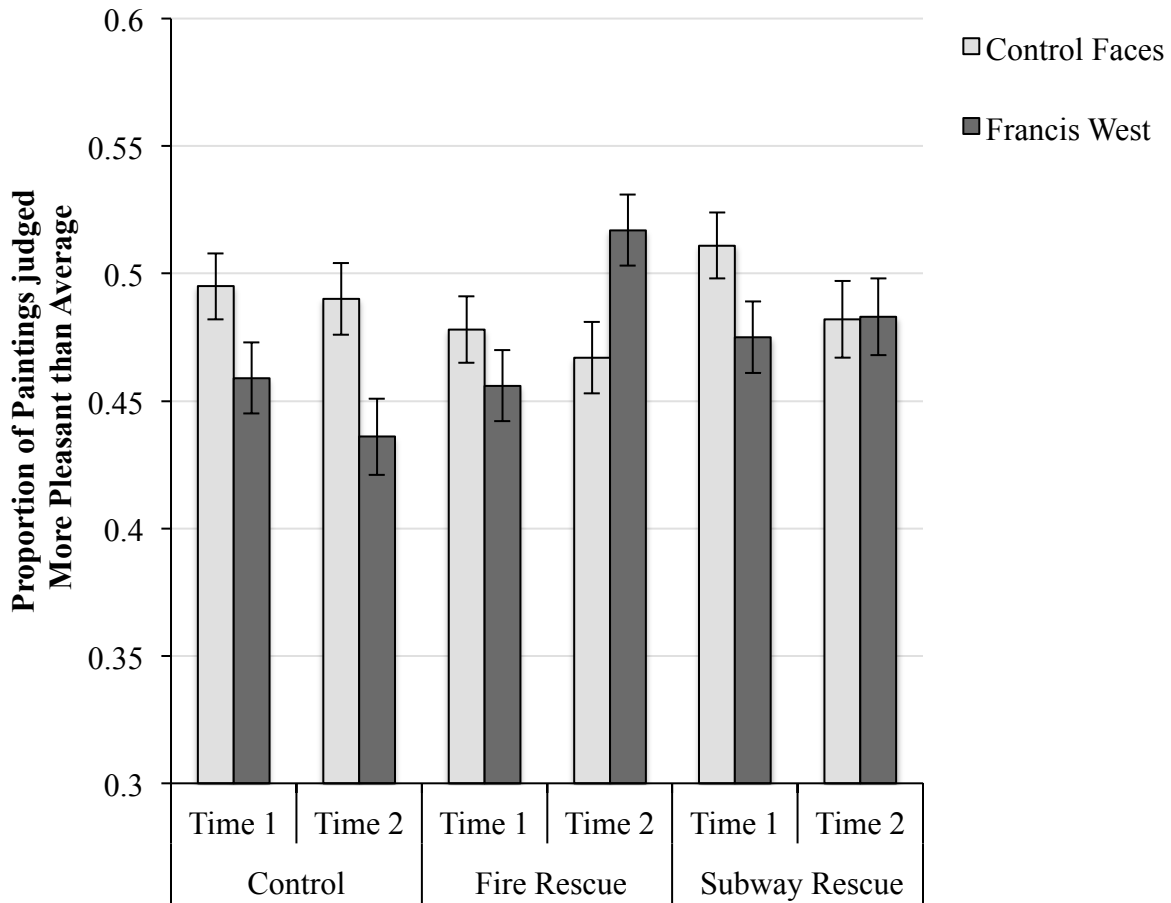


Figure 43. Mean proportion of paintings judged as more pleasant than average in Study 17, by information condition, time, and prime type. Error bars are standard errors.

**Bimodality.** Figure 44 shows the frequency distributions of the Francis West trials at Time 1, and Figure 45 shows the frequency distributions of the Francis West trials within each information condition at Time 2. As was the case in Study 16, the strongly bimodal pattern observed in earlier work with the unmodified AMP was again eliminated here.

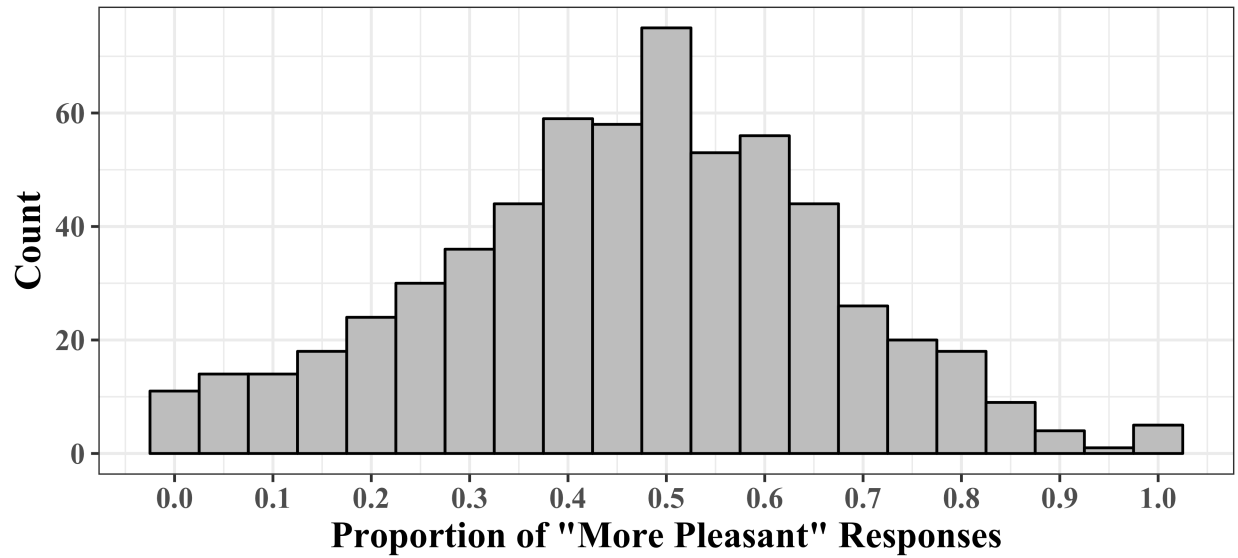


Figure 44. Frequency distribution of the proportion of paintings judged to be more pleasant than average on trials in which Francis West was the prime image at Time 1 in Study 17.

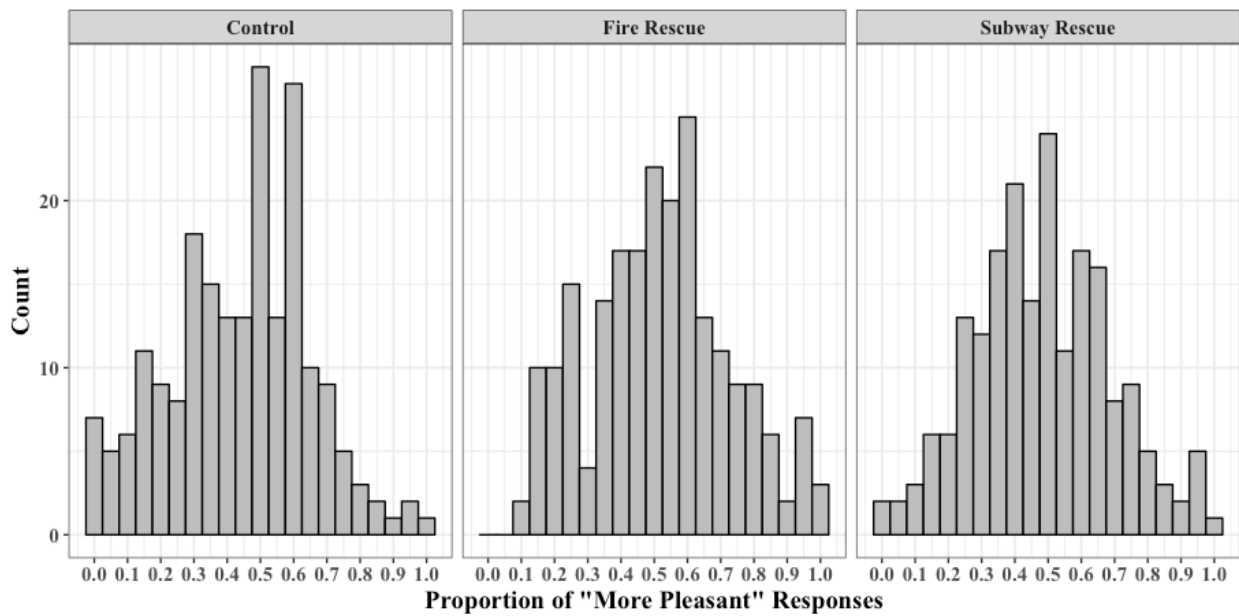


Figure 45. Frequency distributions of the proportion of paintings judged to be more pleasant than average on trials in which Francis West was the prime image in Study 17, by information condition.

**Explicit evaluations.** The pattern of explicit evaluations mirrored that found in Study 16, with a significant interaction between time and information condition on explicit index scores,  $F(2,616) = 957.15, p < .001, \eta_p^2 = .757$ . Explicit evaluations of Francis West were very low (on a scale from 1 – 7, with higher scores indicating more positive evaluations) at Time 1 ( $M = 1.28, SD = .76$ ), remained so at Time 2 in the control condition ( $M = 1.14, SD = .42$ ), and had increased in the subway rescue ( $M = 2.35, SD = 1.37$ ) and fire rescue conditions ( $M = 5.96, SD = 1.31$ ). The increases from Time 1 to Time 2 in the fire rescue and subway rescue conditions, the slight decrease over time in the control condition, and the differences among the three conditions at Time 2 were all significant, all  $ts > 3.4$ , all  $ps \leq .001$ .

## **Discussion**

Revision in implicit impressions of Francis West was once again stronger in the fire rescue (reinterpretation) condition than in the subway rescue (control positive) condition, such that evaluations reversed in the former, becoming significantly positive, while only shifting to neutrality in the latter. Importantly, this difference emerged here in the context of stronger initial formation of a negative initial implicit impression than what was observed in Study 16, with the overall sample now evidencing initial formation.

## **General Discussion**

The additional analyses and experiments presented in Chapter IV take a variety of approaches to assessing the features of automaticity that are exhibited by the revised impressions first studied in Chapters II-III. Of primary focus has been the question of whether responses on the AMP and its variants – the main measure used in this work for its attractive metrics (see Payne & Lundberg, 2014, for a review) and lack of reliance on response time and noise-inducing

block orders – truly reflect an unintentional process of evaluating the primes, or instead reflect intentional rating of the primes. This is an important issue given that unintentionality is often regarded as the most central unifying feature of responses on the family of tasks regarded as “implicit measures” in social cognition (Ferguson, 2007; Ferguson & Fukukura, 2012; Payne et al., 2008; see also discussions in De Houwer et al., 2009; De Houwer & Moors, 2012). For this reason, the implicitness of the AMP has been discussed before (Bar-Anan & Nosek, 2012), with the bulk of the evidence showing that post-hoc reports of intentional responding on the part of participants with the strongest effect sizes is probably due to confabulation (Payne et al., 2013) and that AMP effects persist even when the correlation with self-reported intentional responding is eliminated (Gawronski & Ye, 2015; see also Gawronski, Cunningham, LeBel, & Deutsch, 2010). However, the previously unidentified pattern of bimodal proportions of pleasant responses on the AMP, which was present in most of the studies in this package, seemed to warrant a reexamination of this issue, because such a pattern could be produced by dual processes (Freeman & Dale, 2012). A mixed set of some participants who were intentionally judging the primes and some who were following the task instructions to rate the pictographs constituted one such plausible (though not exclusive) cause of the pattern.

The analysis and studies in this chapter largely support the claims that 1) the bimodal pattern is not (at least entirely) due to intentional rating of the primes and 2) the critical priming effects from earlier studies are not dependent on the presence of the anomalous pattern. The initial analyses show that participants falling into the bimodal tails of the distributions still show evidence of being impacted by the qualities of the pictographs themselves and do not notably differ from other participants in their responses to the various explicit measures included across Studies 1a-9. The bimodal data patterns were also successfully reproduced by the multinomial



modeling approach first used in Study 8, and showed that these trends could be produced by bimodality in the pattern of underlying evaluative responses to the primes without marked bimodality in the likelihood of misattribution, and also suggested that a correlation between the evaluation of the primes and likelihood of misattribution could exacerbate bimodality in the final responses – an unsurprising relationship even if no participants are intentionally judging the primes. It is important to note, however, that the bimodal pattern in the A parameter predicted by this model is currently without clear explanation, especially if it were to be found to occur across different paradigms in the present work (e.g. Francis West, Jonathan and Elizabeth). Finally, the elimination of participants in the extremes of the distributions, though drastic, does not result in the loss of the crucial difference in revision between the fire rescue and subway rescue conditions. This is a key comparison for the argument that the relationship between new and old information, and not just the valence and extremity of the new information, is important in driving revision.

The studies in the chapter bolstered the argument by showing that 1) priming effects (and bimodality) persist even when participants are given an option to skip any trial in which they felt that the prime might impact their responses (Studies 13-14), and were encouraged to use that option (Payne et al., 2013); 2) the priming effect in the fire rescue condition replicated under conditions of distraction and considerable cognitive load (Study 15); and 3) priming effects replicated when the bimodality was eliminated by enhanced warning instructions and the use of new stimuli (Studies 16-17). These studies collectively support the conclusion that the AMP priming effects used throughout Chapters II-III to argue for the role of reinterpretation in driving implicit revision do not depend on the bimodal pattern; this fits with the conclusion of earlier

researchers that the evidence continues to support the implicit status of the AMP (Gawronski & Ye, 2015; Payne et al., 2013).

Is it possible that there is still intentional responding on the AMP? The additional analyses and studies in this chapter cannot rule it out, though use of implicit measures has never rested on the assumption that *all* participants complete them properly. The IAT can be faked if participants know of or discover the best strategy for influencing their scores (e.g., to reduce implicit bias, slow down on the *compatible* block; Fiedler & Bluemke, 2005), as can the evaluative priming task (Teige-Mocigemba & Klauer, 2013; see also Teige-Mocigemba & Klauer, 2008; cf. Degner, 2009). But as Teige-Mocigemba and Klauer (2013) noted in their work on the controllability of evaluative priming, “the present findings will in all likelihood not pose a threat to the validity of evaluative-priming measures in most research situations in which participants are not motivated to fake, not given specific directions on how to do so, and/or in which the measurement purpose is often obscured on purpose.” (p. 654) The key question is whether a *non-trivial* portion of participants deviate from the task instructions under typical usage of the task. The observation of bimodality certainly raised the possibility that this was so, but most of the subsequent studies and analyses suggested that despite the unusual pattern, the pattern is possibly not attributable to intentional responding. As Gawronski and Ye (2015) also noted with regard to correlations between self-reported intentionality and priming effects on the AMP, regardless of the ultimate interpretation of that finding, it is straightforward to simply avoid it by adopting changes to the task. An approximation of the strategy employed by Gawronski and Ye (2015) for eliminating a correlation between reported intentional responding and the priming effect on the AMP was successful, though not feasible to widely implement given its taxing nature (Study 15), but the modifications made in Studies 16-17 – the use of

enhanced warnings and the paintings stimuli – appear to offer a straightforward route for eliminating the bimodal distributions produced by the task.

In general, the issues with one implicit measure can also be compensated for through use of other measures that have their own advantages and drawbacks (convergent validity). In future work, it will be advisable to use a wider range of implicit measures as another source of evidence for the conditions under which implicit updating will vs. will not emerge, especially if such measures sometimes disagree (Study 12; Deutsch & Gawronski, 2009; Deutsch, Kordts-Freudinger, Gawronski, & Strack, 2009). Importantly, the basic revision effect was reproduced in Chapter II with an IAT (Study 1b), though further replications with other measures would be desirable. The new studies in this chapter have focused on further tests of the AMP both because doing so helps to validate the results of the majority of the earlier studies in this package that made use of that task, and because the desirable psychometric properties of the AMP (Bar-Anan & Nosek, 2014; Payne et al., 2005; Payne & Lundberg, 2014) make it attractive for continued use (which also motivated its use in the first place) if a way to correct for bimodal distributions could be found (the goal of Studies 15-17). This work on implicit impressions of novel targets (see also Cone & Ferguson, 2015) has made relatively less use of the IAT and related tasks because its strong block order effect is particularly problematic for work with novel targets. With a relatively small amounts of information yet learned about a novel target, the assigned pairing of the target with either the positive or negative key for the entire first half of the task could result in evaluative *conditioning* as much as evaluative measurement (De Houwer, Thomas, & Baeyens, 2001; Hofmann, De Houwer, Perugini, Baeyens, & Crombez, 2010), as that practiced pairing constitutes a significant portion of the participant's entire history of experience with that target. Another prominent alternative is the evaluative priming task (EPT; Fazio et al., 1986;

Fazio et al., 1995; see Herring et al., 2013, for a review), in which participants categorize unambiguously positive and negative stimuli that are preceded by experimental primes, with effects computed as temporal response facilitation scores for each prime type to positive vs. negative targets. Though this is an oft-used implicit measure (Cameron et al., 2010; Herring et al., 2013) and has attractive features, like allowing for an examination of bivalent priming (Zayas & Shoda, 2015), my colleagues and I have preferred the AMP due to the often relatively low reliability of the EPT (Bar-Anan & Nosek, 2014; Fazio & Olson, 2003), which we previously found to be an issue for detecting even implicit impression *formation* effects with the online samples that we have often used. Replication of some of the effects in the present work in the more controlled lab environment using an EPT is, however, an important additional direction for research.

It must not be overlooked that the studies in Chapter IV also provide new evidence on other automaticity features of the revised implicit evaluations, including whether participants are aware of implicit misattribution on a trial-by-trial basis (Studies 13-14) and the efficiency with which implicit evaluations can be produced under load (Study 15). That participants only infrequently used a readily-available option to skip trials on the AMP (Study 13) and PMP (Study 14), despite showing strong priming effects on those studies, suggests that participants do not generally believe that their responses to the pictographs are the product of the primes until they make post-hoc assessments of their overall task performance after the fact (Payne et al., 2013). Finally, whereas Study 3 (in Chapter II) revealed that processing the propositional meaning of new, reinterpreting information while under load curtails revision, Study 15 offered evidence that revised implicit evaluations continued to be evoked while under cognitive load, albeit with a reduced effect size. This is consistent with the findings of prior studies that

imposed distracting tasks on the AMP and continued to produce priming effects (Gawronski et al., 2010; Gawronski & Ye, 2015). Priming may been more attenuated in the present studies than in other work given that the clearly valenced nature of the smile/frown icons could add noise to the task, and the need to attend to the peripheries of the prime stimulus in order to identify the icons could have drawn attention away from the primes to a degree that diminished the implicit responses that they produced.

## Chapter V. General Discussion

Fiction is rife with tales of characters who seem initially to be kind, friendly, and good, only to be revealed through a clearer understanding of events to be quite the opposite – and vice versa. Be it through the discovery of ulterior motives (e.g., Palpatine from *Star Wars*), the unearthing of critical new information (e.g., Sirius Black from *Harry Potter*), or the overturning of misunderstandings (e.g., Boo Radley from *To Kill a Mockingbird*), many characters in literature, film, and television have ultimately shown true colors opposite from initial expectations. Such updates in how characters should be understood often come at the height of the climax, or at the center of critical plot points, and add greater depth of meaning to the work.

If the prevalence of such twists and turns in fiction attests to a human appetite for exercising and challenging our powers of social inference, the case of Richard Jewell at the 1996 Summer Olympics testifies to the dark side of our drive to understand others, showing how quickly a construal can turn when a compelling (even if erroneous) new interpretation of events comes to light. Such errors notwithstanding, however, the general point that social perceivers seem quite capable of rapidly updating their impressions of other people when seemingly clear, diagnostic, and compelling information is revealed fits with the view of a predictive mind adapted to survive in a complex social world.

The studies in this package provide evidence that even *implicit* responses to other people can be updated to reflect relevant new information, thereby diverging from theories and findings that argue that although it is easy to quickly set aside an erroneous explicit (intentional) first impression, it is much harder and slower to set aside an implicit (unintentional) one (Gawronski et al., 2008; Gregg et al., 2006; McConnell et al., 2008; Rydell & McConnell, 2006; Rydell et al., 2007; Wilson et al., 2000). Chapter II provided a series of studies to show that when a strong

and clear reinterpretation of ostensibly negative behavior is revealed (Francis West broke into his neighbors' homes, but for the purpose of saving children from a fire), implicit evaluations of the character update from negative to positive. The studies examined reinterpretation as the mechanism of change through manipulation of the type of information presented, and through self-report measures of reinterpretation, and established the resource-dependence and durability of updating. Chapter III tested broader conditions of implicit updating by both examining a potentially broader family of revision strategies that include elements of negating an earlier understanding of events and replacing it with new information, and testing the effects of reinterpretation in contexts involving a complex real-world issue, intergroup implicit bias, and gender stereotyping. These studies found diverse evidence for the possibility of implicit updating in numerous domains. Finally, the re-analyses and new studies in Chapter IV tested the implicitness and other operating conditions of the updated impressions examined in the present work, finding evidence in support of the implicitness and robustness of the results.

### **Implicit vs. Explicit Updating**

The studies in this package illustrate that it is possible for initial implicit impressions to be updated quickly, often paralleling the updating of explicit impressions, and through a process of propositional reasoning about the validity of different interpretations of information that has often been thought to generally impact only explicit responses (McConnell & Rydell, 2014; McConnell et al., 2008; Rydell & McConnell, 2006; Strack & Deutsch, 2004). This poses difficulties for theoretical treatments of implicit and explicit cognition that rely on dissociated processes or systems, particularly those that assume that propositional processes can influence only explicit responses. This is not to say, however, that implicit impressions will always update in parallel with explicit impressions, a point first made in the General Discussion of Chapter II

that is worth expanding on here. Studies 11 and 12 in Chapter III found some persistent evidence of implicit race bias and implicit gender stereotyping (on the IAT), respectively, and many lines of work have found dissociations between implicit and explicit responses (e.g., Gawronski & LeBel, 2008; Gawronski & Strack, 2004; Gregg et al., 2006; Ranganath & Nosek, 2008; Ratliff & Nosek, 2011; Rydell & McConnell, 2006; Rydell et al., 2006, 2007; see also Gawronski & Bodenhausen, 2006, 2011; Nosek, 2007). Though some differences between measures are possibly due to the structural features of the tasks rather than due to baked-in differences between implicit and explicit cognition, implicit and explicit measures have still differed even when made structurally similar (Payne et al., 2008; replicated by Open Science Collaboration, 2015). Although dual-process theories have suggested that such differences are due to the selective influence of motivation, propositional knowledge, and reasoning on explicit cognition, the present results and many other lines of work show how such factors can influence implicit cognition as well (De Houwer, 2014; Ferguson & Wojnowicz, 2011). To what can dissociations between implicit and explicit impressions then be attributed?

A possible answer is provided by views of implicit and explicit evaluations as emerging from one common, iterative, dynamical process of constructing a response on-line as constraint satisfaction operates continuously over time across diverse networks in the brain (Cunningham, Zelazo, Packer, & Van Bavel, 2007; Ehret, Monroe, & Read, 2015; Ferguson & Wojnowicz, 2011; Monroe & Read, 2008; Van Bavel, Xiao, & Cunningham, 2012; Wojnowicz, Ferguson, Dale, & Spivey, 2009). This perspective suggests that both intentional and unintentional (explicit and implicit) responses are generated in a fundamentally similar way, in that both types of responses are constructed in time and can be influenced by inputs from external and internal context, goals, emotions, propositional knowledge, and memory representations. Differences in



the *inputs* into this constructive process that tend to differ between intentional and unintentional responses – the defining difference between implicit and explicit responses – would then be the source of dissociations.

How might inputs into the construction of a response differ between implicit and explicit responses? The degree to which participants believe in the moment that they are intentionally providing their opinion of a stimulus (vs. doing something else, like judging a different stimulus, or categorizing it) may not delineate inherently different types of cognitive processes. After all, research suggests that subjective intentionality is an inference on the part of an actor that an action was produced by a particular goal (Aarts, Custers, & Marien, 2009; Aarts & van den Bos, 2011; Wegner, 2003; Wegner, Sparrow, & Winerman, 2004; Wegner & Wheatley, 1999; see also De Houwer & Moors, 2012; Moors & De Houwer, 2006), and because introspective insight into mental processes is imperfect (Morsella, 2005; Nisbett & Wilson, 1977; Wilson, 2002), goals can influence cognition even without awareness (e.g., Bargh, Gollwitzer, Lee-Chai, Barndollar, & Trötschel, 2001; Chartrand & Bargh, 1996; Custers & Aarts, 2010; Ferguson, 2008; Kleiman & Hassin, 2013; Marien et al., 2012). Instead, when a participant has a subjective intention to evaluate a stimulus, this intention may bring to bear a different set of inputs into the construction of an evaluation than would otherwise arise, not because such inputs *can't* affect unintentional responses, but rather because the context of intentionally judging a stimulus tends to activate different goals, knowledge, and representations that have come to be associated with that context. In particular, the context of giving an intentional judgment of a stimulus may tend to activate considerations related to social interaction and self-presentation – goals that *can* modulate implicit cognition when sufficiently active, even when externally imposed (e.g., Castelli & Tomelleri, 2008; Mendoza, Gollwitzer, & Amodio, 2010; Sinclair, Lowery, Hardin, &

Colangelo, 2005; Stewart & Payne, 2008; see Ferguson & Wojnowicz, 2011), especially if strong internal motivations are activated (Legault, Gutsell, & Inzlicht, 2011). However, absent particularly strong activation, social goals may tend to be less effective on a trial-by-trial basis during an implicit measure when the task context switches from giving an intentional response to implementing the proximal task instructions. Even if participants are generally aware of the “true” purpose of an implicit measure, the rapid task-switching in working memory that is necessary to follow the focal task (such as categorizing stimuli on the IAT; e.g., Klauer, Schmitz, Teige-Mocigemba, & Voss, 2010; see also Kiesel, Kunde, & Hoffmann, 2007; Monsell, 2003) might suggest that interpersonal goals that are most strongly active during intentional responding may often not dominate in shaping an unintentional response on a trial-by-trial basis to the same degree. Also in support of this account, it is possible for explicit responses to align with implicit responses when the usual goals related to intentional expression, like self-presentation, are reduced by asking participants to rely more on intuitions (e.g., Loersch, McCaslin, & Petty, 2011).

Under the framework above, explicit responses may change more quickly than implicit responses when new information is learned in a manner that associates it primarily with a goal to report that information intentionally. For instance, when minimal counter-stereotypical information is learned about Jonathan and Elizabeth (Studies 12 and 14; Cao & Banaji, 2016), participants clearly encode the information – after all, they are able to report it later. Why then is this individuating information not dominant on the IAT? One possibility is that participants learn it as something to *report* later, and therefore learn it most strongly under an “expression” goal rather than more generally. After all, they know it to be information they need for this experiment alone, rather than after they complete the study. The frequent task-set switching that

occurs during the IAT (Klauer et al., 2010) may make it particularly unlikely that this expression-context-dependent knowledge will be brought to bear on IAT responses. Some of the moderators of implicit revision discussed in the preceding chapters, like immersion, realism, extremity, motivation, and convincingness of new information may produce fast implicit revision by discouraging the encoding of the new information in a way contextualized to facilitate reporting, and instead encouraging more general learning, a direction that future research might explore. Indeed, Brannon and Gawronski (2017) recently showed in a replication and extension of the Francis West work (and the work presented by Cone & Ferguson, 2015) that the updated implicit impressions were not restricted to the context in which the new information was learned, in contrast to prior work showing that more moderately valenced new information often produces context-specific implicit responses (Gawronski et al., 2010; Rydell & Gawronski, 2009; see Gawronski & Cesario, 2013).

### **Revision in the World**

It is possible, of course, that new information might be seen as highly relevant, generally important, and believable, and yet still fail to override a person's earlier implicit responses even when he or she sees it as more widely applicable than "just" for reporting purposes. The experience of being disgusted with one's own implicit group-based biases, and expressing a wish to change them, has been frequently observed (Banaji & Greenwald, 2013), and recent findings have reiterated that implicit impressions of known groups are difficult to durably revise (Lai et al., 2016; see also Forscher et al., 2017; Bargh, 1999). In such cases, people certainly link egalitarian information with goals of intentional expression, in that they definitely report egalitarian ideals when asked for their views, but they might also wish to integrate their egalitarian beliefs into their implicit cognition with less success (though see Moskowitz,

Gollwitzer, Wasel, & Schaal, 1999; Moskowitz & Ignarri, 2009; Moskowitz & Li, 2011; Moskowitz, Salomon, & Taylor, 2000). Especially with groups about which a person has learned over the course of a lifetime, it may be particularly difficult to isolate, identify, and activate all relevant older representations that might become active and drive implicit responses in order to update them in light of the new information, especially early in the construction of a response before egalitarian goals come online to constrain the produced evaluation.

There are a number of reasons why new information may not fully integrate with all relevant memories (of highly familiar and well-learned groups) that could be activated on an implicit measure when unconstrained by intentional response goals. One, which has already been extensively discussed in Chapter III, is that new information is more likely to be incorporated into or override an earlier memory if those earlier memories are retrieved first (Hardt et al., 2010; Hubbach, 2011; Lee et al., in press), but tends to be encoded as a separate trace otherwise, allowing for either the initial or new response to be retrieved in the future depending on context (Gawronski & Cesario, 2013; Bouton, 1994). In the case of groups, context can lead to the activation of different group subtypes that are associated with different implicit impressions (e.g., Maddux, Barden, Brewer, & Petty, 2005). To the extent that the relevant memories underlying an implicit response are less likely to be retrieved over time due to differences in context, forgetting, or the simple introspective difficulty of identifying the true reasons behind our thoughts (Wilson et al., 1989), such updating may be more difficult (Gershman et al., 2017). Yet, theoretical discussion of successful reconsolidation in the context of clinical treatment may offer productive future directions for work on implicit impression change to explore (Lane et al., 2015).

A recent intervention designed to durably reduce explicit transphobia (on relatively well-disguised questionnaires administered over multiple months) by first inducing participants to generate introspective reasons for their discomfort with transgender people before presenting them with new perspectives on experiencing the world as transgender showed strong evidence of revised explicit impressions (Broockman & Kalla, 2016). Their procedure seemed to draw on many elements of the techniques highlighted by Lane and colleagues (2015) as uniting successful yet diverse clinical traditions that may all draw on reconsolidation update mechanisms. These techniques have notable similarities to the negate-plus-affirm/subtract-plus-add approach of reinterpretation, and so it remains possible that efforts to revise even *implicit* impressions of entire groups that make use of techniques along these lines could be effective. For example, future work could ask participants to introspect on their intuitive, spontaneous feelings toward Muslims, the elderly, or Black people, and to try to identify thoughts that come to mind that might contribute to those reactions. Along the lines of Broockman and Kalla (2016), potential reasons could even be presented to participants to further aid in memory retrieval (such as stereotypes, prominent events, and beliefs of national figures to which they may have been exposed). Then, ways to reframe those ideas from the perspective of the stigmatized group could be presented. A series of studies could aim to isolate exactly which ingredients in the kitchen-sink strategy employed by Broockman and Kalla (2016) might generalize to successful implicit impression change, making use as well of insights from basic research on reconsolidation, such as the importance of expectancy violation during retrieval (Sevenster et al., 2012, 2013, 2014).

Of course, updating with known groups presents more challenges than simply retrieving the relevant memories to revise; people often learn about individual *members* of groups rather than the groups as a whole, and this distinction allows for impressions of the overall groups to be

“protected” by individuation and subtyping (e.g., Kunda & Oleson, 1995, 1997), especially if the perceiver is motivated to maintain the first impression. Nonetheless, as the study on big-game hunting (Study 10) suggests, to the degree that a person *does* find new information about a group persuasive and important, implicit impressions of well-known groups can change, even in complicated applied settings in which the new information is not as clean and clear-cut as in the Francis West paradigm. I thus remain optimistic that in future work examining applications of the findings described here, routes for the updating of implicit impressions of known groups will be identified and prove fruitful, especially with the leads provided by the studies found in Chapter III and recent advances in memory updating.

## **Conclusion**

It is possible for implicit first impressions to be updated through reinterpretation, bringing even our spontaneous responses to other people into harmony with what we learn to be true. Though there is still much work to be done to understand the full scope, implications, and limitations of the routes to revision examined here, the endeavor ultimately promises to paint a more complete picture of how the mind navigates the complex, ever-changing social world. Instead of being fundamentally incapable of fully aligning with important new truths, we are at times able to adapt to those truths at both an explicit and implicit level.

## REFERENCES

- Aarts, H., Custers, R., & Marien, H. (2009). Priming and authorship ascription: When nonconscious goals turn into conscious experiences of self-agency. *Journal of Personality and Social Psychology*, 96(5), 967–979. <http://doi.org/10.1037/a0015000>
- Aarts, H., & van den Bos, K. (2011). On the foundations of beliefs in free will: intentional binding and unconscious priming in self-agency. *Psychological Science*, 22(4), 532–537. <http://doi.org/10.1177/0956797611399294>
- Alberini, C. M. (2011). The role of reconsolidation and the dynamic process of long-term memory formation and storage. *Frontiers in Behavioral Neuroscience*. <http://doi.org/10.3389/fnbeh.2011.00012/abstract>
- Amodio, D. M., & Ratner, K. G. (2011). A memory systems model of implicit social cognition. *Current Directions in Psychological Science*, 20(3), 143–148. <http://doi.org/10.1177/0963721411408562>
- Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2014). The rules of implicit evaluation by race, religion, and age. *Psychological Science*, 25(9), 1804–1815. <http://doi.org/10.1177/0956797614543801>
- Banaji, M. R. & Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. Random House.
- Bar, M. (2011). The proactive brain. In M. Bar (Ed.), *Predictions in the brain: Using our past to generate a future* (pp. 13–26). New York, NY: Oxford University Press.
- Bar-Anan, Y., & Nosek, B. A. (2012). Reporting intentional rating of the primes predicts priming effects in the affective misattribution procedure. *Personality and Social Psychology Bulletin*, 38(9), 1194–1208. <http://doi.org/10.1177/0146167212446835>

- Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods*, 46(3), 668–688. <http://doi.org/10.3758/s13428-013-0410-6>
- Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (2nd ed., pp. 1-40). Hillsdale, NJ: Erlbaum.
- Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 361-382). New York: Guilford Press.
- Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The automated will: nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology*, 81(6), 1014–1027.
- Baron, A. S., & Banaji, M.R. (2006). The development of implicit attitudes: evidence of race evaluations from ages 6 and 10 and adulthood. *Psychological Science*, 17, 53–58. doi: 10.1111/j.1467-9280.2005.01664.x
- Bartlett, F. A. (1932). *Remembering: A study in experimental social psychology*. New York: Cambridge University Press.
- Barrett, L. F. (2012). Emotions are real. *Emotion*, 12(3), 413–429. <http://doi.org/10.1037/a0027555>
- Barrett, L. F., & Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in Experimental Social Psychology* (1st ed., Vol. 41, pp. 167–218). Elsevier Inc. [http://doi.org/10.1016/S0065-2601\(08\)00404-8](http://doi.org/10.1016/S0065-2601(08)00404-8)



- Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6(1), 57–86.
- Bijleveld, E., Custers, R., & Aarts, H. (2009). The unconscious eye opener: Pupil dilation reveals strategic recruitment of resources upon presentation of subliminal reward cues. *Psychological Science*, 20(11), 1313–1315. <http://doi.org/10.1111/j.1467-9280.2009.02443.x>
- Blechert, J., Sheppes, G., Di Tella, C., Williams, H., & Gross, J. J. (2012). See what you think: Reappraisal modulates behavioral and neural responses to social stimuli. *Psychological Science*, 23(4), 346–353. <http://doi.org/10.1177/0956797612438559>
- Brannon, S. M., & Gawronski, B. (2017). A second chance for first impressions? Exploring the context-(in)dependent updating of implicit evaluations. *Social Psychological and Personality Science*, 8(3), 275–283. <http://doi.org/10.1177/1948550616673875>
- Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, 352(6282), 220–224. <http://doi.org/10.1126/science.aad8127>
- Brooks, A. W. (2014). Get excited: Reappraising pre-performance anxiety as excitement. *Journal of Experimental Psychology: General*, 143(3), 1144–1158. <http://doi.org/10.1037/a0035325>
- Boucher, K. L., & Rydell, R. J. (2012). Impact of negation salience and cognitive resources on negation during evaluation formation. *Personality and Social Psychology Bulletin*, 38, 1329-1342.
- Bouton, M. E. (1994). Context, ambiguity, and classical conditioning. *Current Directions in Psychological Science*, 3(2), 49–53. <http://doi.org/10.1111/1467-8721.ep10769943>

- Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1999). The affect system has parallel and integrative processing components: Form follows function. *Journal of Personality and Social Psychology*, 76(5), 839-855.
- Cameron, C. D., Brown-Iannuzzi, J. L., & Payne, B. K. (2012). Sequential priming measures of implicit social cognition: A meta-analysis of associations with behavior and explicit attitudes. *Personality and Social Psychology Review*, 16(4), 330–350.  
<http://doi.org/10.1177/1088868312440047>
- Cao, J., & Banaji, M. R. (2016). The base rate principle and the fairness principle in social judgment. *Proceedings of the National Academy of Sciences*, 113(27), 7475–7480.  
<http://doi.org/10.1073/pnas.1524268113>
- Castelli, L., & Tomelleri, S. (2008). Contextual effects on prejudiced attitudes: When the presence of others leads to more egalitarian responses. *Journal of Experimental Social Psychology*, 44(3), 679–686. <http://doi.org/10.1016/j.jesp.2007.04.006>
- Chartrand, T. L., & Bargh, J. A. (1996). Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. *Journal of Personality and Social Psychology*, 71(3), 464–478.
- Chen, M., & Bargh, J. A. (1997). Nonconscious behavioral confirmation processes: The self-fulfilling consequences of automatic stereotype activation. *Journal of Experimental Social Psychology*, 33, 541–560.
- Chester, D. S., & DeWall, C. N. (2017). Combating the sting of rejection with the pleasure of revenge: A new look at how emotion shapes aggression. *Journal of Personality and Social Psychology*, 112(3), 413–430. <http://doi.org/10.1037/pspi0000080>

- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36, 181–253.
- Cone, J., & Ferguson, M. J. (2015). He did *what*? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, 108, 37-57.  
<http://doi.org/10.1037/pspa0000014>
- Cone, J., Mann, T. C., & Ferguson, M. J. (in press). Changing our implicit minds: How, when, and why implicit evaluations can be rapidly revised. *Advances in Experimental Social Psychology*. Elsevier Inc. <http://doi.org/10.1016/bs.aesp.2017.03.001>
- Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology*, 89(4), 469–487.
- Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, 83(6), 1314–1329. <http://doi.org/10.1037//0022-3514.83.6.1314>
- Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of black and white faces. *Psychological Science*, 15(12), 806–813.
- Cunningham, W. A., Zelazo, P. D., & Packer, D. J. (2007). The iterative reprocessing model: A multilevel framework for attitudes and evaluation. *Social Cognition*, 25(5), 736–760.
- Custers, R., & Aarts, H. (2010). The unconscious will: How the pursuit of goals operates outside of conscious awareness. *Science*, 329(5987), 47–50.  
<http://doi.org/10.1126/science.1188595>

- Darley, J. M., & Fazio, R. H. (1980). Expectancy confirmation processes arising in the social interaction sequence. *American Psychologist*, 35, 867–881. <http://dx.doi.org/10.1037/0003-066X.35.10.867>.
- Degner, J. (2009). On the (un-)controllability of affective priming: Strategic manipulation is feasible but can possibly be prevented. *Cognition and Emotion*, 23, 327–354.
- Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts*. New York, NY: Viking.
- De Houwer, J. (2001). A Structural and process analysis of the implicit association test. *Journal of Experimental Social Psychology*, 37(6), 443–451.  
<http://doi.org/10.1006/jesp.2000.1464>
- De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, 8(7), 342–353.
- De Houwer, J., & Moors, A. (2012). How to define and examine implicit processes. In R. Proctor & J. Capaldi (Eds.), *Psychology of science: Implicit and explicit processes* (pp. 183–198). New York, NY: Oxford.
- De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin*, 135(3), 347–368.  
<http://doi.org/10.1037/a0014211>
- De Houwer, J., Thomas, S., & Baeyens, F. (2001). Associative learning of likes and dislikes: A review of 25 years of research on human evaluative conditioning. *Psychological Bulletin*, 127(6), 853–869.
- Deutsch, R., & Gawronski, B. (2009). When the method makes a difference: Antagonistic effects on “automatic evaluations” as a function of task characteristics of the measure. *Journal*

- of Experimental Social Psychology*, 45(1), 101–114.  
<http://doi.org/10.1016/j.jesp.2008.09.001>
- Deutsch, R., Kordts-Freudinger, R., Gawronski, B., & Strack, F. (2009). Fast and Fragile. *Experimental Psychology (Formerly "Zeitschrift Für Experimentelle Psychologie")*, 56(6), 434–446. <http://doi.org/10.1027/1618-3169.56.6.434>
- Dolderer, M., Mummendey, A., & Rothermund, K. (2009). And yet they move: The impact of direction of deviance on stereotype change. *Personality and Social Psychology Bulletin*, 35, 1368-1381.
- Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science*, 11, 319-323.
- Eberhardt, J.L., Goff, P.A., Purdie, V.J., & Davies, P.G. (2004). Seeing Black: Race, crime, and visual processing. *Journal of Personality and Social Psychology*, 87, 876–893.
- Edelman, S. (2008). *Computing the mind: How the mind really works*. Oxford: Oxford University Press.
- Ehret, P. J., Monroe, B. M., & Read, S. J. (2015). Modeling the dynamics of evaluation: A multilevel neural network implementation of the iterative reprocessing model. *Personality and Social Psychology Review*, 19(2), 148–176.  
<http://doi.org/10.1177/1088868314544221>
- Eiser, J. R., Fazio, R. H., Stafford, T., & Prescott, T. J. (2003). Connectionist simulation of attitude learning: Asymmetries in the acquisition of positive and negative evaluations. *Personality and Social Psychology Bulletin* 29, 1221-1235.
- Eiser, J. R., Stafford, T., & Fazio, R. H. (2008). Expectancy-confirmation in attitude learning: A connectionist account. *European Journal of Social Psychology*, 38, 1023-1032.

- Fazio, R. H. (2007). Evaluations as object-evaluation associations of varying strength. *Social Cognition, 25*, 603–637.
- Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology, 87*(3), 293–311.  
<http://doi.org/10.1037/0022-3514.87.3.293>
- Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology, 69*, 1013–1027. doi:10.1037/0022-3514.69.6.1013
- Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual Review of Psychology, 54*(1), 297–327.  
<http://doi.org/10.1146/annurev.psych.54.101601.145225>
- Fazio, R. H., Pietri, E. S., Rocklage, M. D., & Shook, N. J. (2015). Positive versus negative valence: Asymmetries in attitude formation and generalization as fundamental individual differences. In J. M. Olson & M. P. Zanna (Eds.), *Advances in Experimental Social Psychology* (Vol. 51, pp. 97-146). Burlington: Academic Press.
- Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*, 229–238.  
doi:10.1037/0022-3514.50.2.229
- Ferguson, M. J. (2007). The automaticity of evaluation. Invited chapter in J. A. Bargh (Ed.), *Social Psychology and the Unconscious: The automaticity of higher mental processes* (pp. 219-264). Psychology Press.

- Ferguson, M. J. (2008). On becoming ready to pursue a goal you don't know you have: Effects of nonconscious goals on evaluative readiness. *Journal of Personality and Social Psychology*, 95(6), 1268–1294. <http://doi.org/10.1037/a0013263>
- Ferguson, M. J., Bargh, J. A., & Nayak, D. A. (2005). After-affects: How automatic evaluations influence the interpretation of subsequent, unrelated stimuli. *Journal of Experimental Social Psychology*, 41(2), 182–191. <http://doi.org/10.1016/j.jesp.2004.05.008>
- Ferguson, M. J., & Furr, R. M. (2012). Likes and dislikes: A social cognitive perspective. In S. Fiske, & C. N. Macrae (Eds.), *Sage Handbook of Social Cognition* (pp. 165-189). Los Angeles: SAGE.
- Ferguson, M. J., Mann, T. C., & Wojnowicz, M. (2014). Rethinking duality: Criticisms and ways forward. In J. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual process theories of the social mind* (pp. 578-594). Guilford Press.
- Ferguson, M. J., & Wojnowicz, M. T. (2011). The when and how of evaluative readiness: A social cognitive neuroscience perspective. *Social and Personality Psychology Compass*, 5(12), 1018–1038.
- Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the implicit association tests. *Basic and Applied Social Psychology*, 27(4), 307–316. [http://doi.org/10.1207/s15324834basp2704\\_3](http://doi.org/10.1207/s15324834basp2704_3)
- Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 23, pp. 1–74). San Diego, CA: Academic Press.

- Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2017). *A meta-analysis of change in implicit bias*. Unpublished manuscript.
- Freeman, J. B., & Dale, R. (2012). Assessing bimodality to detect the presence of a dual cognitive process. *Behavior Research Methods*, 45(1), 83–97.  
<http://doi.org/10.3758/s13428-012-0225-x>
- Fujita, K., & Carnevale, J. J. (2012). Transcending temptation through abstraction: The role of construal level in self-control. *Current Directions in Psychological Science*, 21(4), 248–252. <http://doi.org/10.1177/0963721412449169>
- Gawronski, B., & Ye, Y. (2014). What drives priming effects in the affect misattribution procedure? *Personality and Social Psychology Bulletin*, 40(1), 3–15.  
<http://doi.org/10.1177/0146167213502548>
- Gawronski, B., & Ye, Y. (2015). Prevention of intention invention in the affect misattribution procedure. *Social Psychological and Personality Science*, 6(1), 101–108.  
<http://doi.org/10.1177/1948550614543029>
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit evaluation change. *Psychological Bulletin*, 132, 692–731.
- Gawronski, B., & Bodenhausen, G. V. (2011). The associative–propositional evaluation model: Theory, evidence, and open questions. *Advances in Experimental Social Psychology* (1st ed., Vol. 44, pp. 59–127). Elsevier Inc. <http://doi.org/10.1016/B978-0-12-385522-0.00002-0>
- Gawronski, B., & Bodenhausen, G. V. (2014). The associative-propositional evaluation model: Operating principles and operating conditions of evaluation. In J. W. Sherman, B.



- Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 188-203). New York, NY: Guilford Press.
- Gawronski, B., & Cesario, J. (2013). Of mice and men: What animal research can tell us about context effects on automatic responses in humans. *Personality and Social Psychology Review, 17*(2), 187–215.
- Gawronski, B., Cunningham, W. A., LeBel, E. P., & Deutsch, R. (2010). Attentional influences on affective priming: Does categorisation influence spontaneous evaluations of multiply categorisable objects? *Cognition & Emotion, 24*(6), 1008–1025.  
<http://doi.org/10.1080/02699930903112712>
- Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When “Just Say No” is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology, 44*(2), 370–377.  
<http://doi.org/10.1016/j.jesp.2006.12.004>
- Gawronski, B., & LeBel, E. P. (2008). Understanding patterns of attitude change: When implicit measures show change, but explicit measures do not. *Journal of Experimental Social Psychology, 44*(5), 1355–1361. <http://doi.org/10.1016/j.jesp.2008.04.005>
- Gawronski, B., Rydell, R. J., Vervliet, B., & De Houwer, J. (2010). Generalization versus contextualization in automatic evaluation. *Journal of Experimental Psychology: General, 139*(4), 683–701.
- Gawronski, B., & Strack, F. (2004). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology, 40*(4), 535–542. <http://doi.org/10.1016/j.jesp.2003.10.005>

- Gershman, S. J., Monfils, M. H., Norman, K. A., & Niv, Y. (2017). The computational nature of memory modification. *eLife*, 6:e23763. <http://doi.org/10.7554/eLife.23763.001>
- Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology*, 60(4), 509-517.
- Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5), 1029–1046. <http://doi.org/10.1037/a0015141>
- Greenwald, A. G. (1968). Cognitive learning, cognitive response to persuasion, and evaluation change. In A. G. Greenwald, T. C. Brock, and T. M. Ostrom (Eds.), *Psychological foundations of evaluations* (pp. 147-170). New York: Academic Press.
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102, 4-27.
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74, 1464–1480.
- Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, 85(2), 197–216. <http://doi.org/10.1037/0022-3514.85.2.197>
- Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, 90(1), 1–20.
- Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, 290, 181–197.

- Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, 143(3), 1369–1392.  
<http://doi.org/10.1037/a0035028>
- Hardt, O., Einarsson, E. Ö., & Nader, K. (2010). A bridge over troubled water: Reconsolidation as a link between cognitive and neuroscientific memory research traditions. *Annual Review of Psychology*, 61(1), 141–167.  
<http://doi.org/10.1146/annurev.psych.093008.100455>
- Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *The Annals of Statistics*, 13, 70–84.
- Hassin, R. R. (2013). Yes it can: On the functional abilities of the human unconscious. *Perspectives on Psychological Science*, 8(2), 195–207.
- Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis*. New York: Guilford Press.
- Helman, E., Stoller, R. M., & Freeman, J. B. (2015). Advanced mouse-tracking analytic techniques for enhancing psychological science. *Group Processes & Intergroup Relations*, 18(3), 384–401. <http://doi.org/10.1177/1368430214538325>
- Helmholtz, H. V. (1860). *Treatise on physiological optics*. New York: Dover.
- Herring, D. R., White, K. R., Jabeen, L. N., & Hinojos, M. (2013). On the automatic activation of attitudes: a quarter century of evaluative priming research. *Psychological Bulletin*, 139(5), 1062–1089. <http://doi.org/10.1037/a0031309>
- Higgins, E. T., Rholes, W. S., & Jones, C. R. (1977). Category accessibility and impression formation. *Journal of Experimental Social Psychology*, 13(2), 141–154.  
[http://doi.org/10.1016/S0022-1031\(77\)80007-3](http://doi.org/10.1016/S0022-1031(77)80007-3)

- Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, 136(3), 390–421.  
<http://doi.org/10.1037/a0018916>
- Horcajo, J., Briñol, P., & Petty, R. E. (2010). Consumer persuasion: Indirect change and implicit balance. *Psychology and Marketing*, 27(10), 938–963. <http://doi.org/10.1002/mar.20367>
- Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, 59, 21–47.
- Huang, Y., & Rao, R. P. N. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(5), 580–593. <http://doi.org/10.1002/wcs.142>
- Hugenberg, K., & Bodenhausen, G. V. (2003). Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science*, 14(6), 640–643.
- Hugenberg, K., & Bodenhausen, G. V. (2004). Ambiguity in social categorization: The role of prejudice and facial affect in race categorization. *Psychological Science*, 15(5), 342–345.
- Hughes, S., Barnes-Holmes, D., & De Houwer, J. (2011). The dominance of associative theorizing in implicit attitude research: Propositional and behavioral alternatives. *The Psychological Record*, 61(3), 465–496.
- Hupbach, A. (2011). The specific outcomes of reactivation-induced memory changes depend on the degree of competition between old and new information. *Frontiers in Behavioral Neuroscience*, 5. <http://doi.org/10.3389/fnbeh.2011.00033/full>
- Hutchings, P. B., & Haddock, G. (2008). Look Black in anger: The role of implicit prejudice in the categorization and perceived emotional intensity of racially ambiguous faces. *Journal of Experimental Social Psychology*, 44(5), 1418–1420.  
<http://doi.org/10.1016/j.jesp.2008.05.002>

- Imhoff, R., Schmidt, A. F., Bernhardt, J., Dierksmeier, A., & Banse, R. (2011). An inkblot for sexual preference: A semantic variant of the Affect Misattribution Procedure. *Cognition & Emotion*, 25(4), 676–690. <http://doi.org/10.1080/02699931.2010.508260>
- Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 513–541.
- Johnston, L., & Hewstone, M. (1992). Cognitive models of stereotype change: III. Subtyping and the perceived typicality of disconfirming group members. *Journal of Experimental Social Psychology*, 28, 360-386.
- Johnson, I. R., Kopp, B. M., & Petty, R. E. (in press). Just say no! (and mean it): Meaningful negation as a tool to modify automatic racial attitudes. *Group Processes & Intergroup Relations*. <http://doi.org/10.1177/1368430216647189>
- Kiesel, A., Kunde, W., & Hoffmann, J. (2007). Unconscious priming according to multiple S-R rules. *Cognition*, 104(1), 89–105. <http://doi.org/10.1016/j.cognition.2006.05.008>
- Kiesel, A., Kunde, W., Pohl, C., Berner, M. P., & Hoffmann, J. (2009). Playing chess unconsciously. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(1), 292–298. <http://doi.org/10.1037/a0014499>
- Klauer, K. C., & Musch, J. (2002). Goal-dependent and goal-independent effects of irrelevant evaluations. *Personality and Social Psychology Bulletin*, 28(6), 802–814.
- Klauer, K. C., Schmitz, F., Teige-Mocigemba, S., & Voss, A. (2010). Understanding the role of executive control in the Implicit Association Test: Why flexible people have small IAT effects. *The Quarterly Journal of Experimental Psychology*, 63(3), 595–619. <http://doi.org/10.1080/17470210903076826>

- Kleiman, T., & Hassin, R. R. (2013). When conflicts are good: Nonconscious goal conflicts reduce confirmatory thinking. *Journal of Personality and Social Psychology*, 105(3), 374–387. <http://doi.org/10.1037/a0033608>
- Koster, E. H. W., Crombez, G., Van Damme, S., Verschuere, B., & De Houwer, J. (2004). Does Imminent Threat Capture and Hold Attention? *Emotion*, 4(3), 312–317. <http://doi.org/10.1037/1528-3542.4.3.312>
- Kreps, T. A., & Monin, B. (2014). Core values versus common sense: Consequentialist views appear less rooted in morality. *Personality and Social Psychology Bulletin*, 40(11), 1529–1542. <http://doi.org/10.1177/0146167214551154>
- Krieglmeyer, R., & Sherman, J. W. (2012). Disentangling stereotype activation and stereotype application in the stereotype misperception task. *Journal of Personality and Social Psychology*, 103(2), 205–224. <http://doi.org/10.1037/a0028764>
- Kunda, Z., & Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: constructing grounds for subtyping deviants. *Journal of Personality and Social Psychology*, 68(4), 565–579.
- Kunda, Z., & Oleson, K. C. (1997). When exceptions prove the rule: how extremity of deviance determines the impact of deviant examples on stereotypes. *Journal of Personality and Social Psychology*, 72(5), 965–979.
- Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., et al. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, 143(4), 1765–1785. <http://doi.org/10.1037/a0036260>

- Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., et al. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, 145(8), 1001–1016.  
<http://doi.org/10.1037/xge0000179>
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4(863), 1–12.  
<http://doi.org/10.3389/fpsyg.2013.00863/abstract>
- Lane, R. D., Ryan, L., & Nadel, L. (2015). Memory reconsolidation, emotional arousal and the process of change in psychotherapy: New insights from brain science. *Behavioral and Brain Sciences*, 38, 1–64.
- Lavandera, E. (2015, May 21). Texas hunter bags his rhino on controversial hunt in Namibia. *CNN*. Retrieved from <http://www.cnn.com/2015/05/19/africa/namibia-rhino-hunt/index.html>
- Lee, J. L. C., Nader, K., & Schiller, D. (in press). An update on memory reconsolidation updating. *Trends in Cognitive Sciences*. <http://doi.org/10.1016/j.tics.2017.04.006>
- Lee, J. L. C. (2009). Reconsolidation: maintaining memory relevance. *Trends in Neurosciences*, 32(8), 413–420. <http://doi.org/10.1016/j.tins.2009.05.002>
- Legault, L., Gutsell, J. N., & Inzlicht, M. (2011). Ironic effects of antiprejudice messages: How motivational interventions can reduce (but also increase) prejudice. *Psychological Science*, 22(12), 1472–1477. <http://doi.org/10.1177/0956797611427918>
- Loersch, C., Mccaslin, M. J., & Petty, R. E. (2011). Exploring the impact of social judgeability concerns on the interplay of associative and deliberative attitude processes. *Journal of Experimental Social Psychology*, 47(5), 1029–1032.

- Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago Face Database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 47, 1122-1135.
- Maddux, W. W., Barden, J., Brewer, M. B., & Petty, R. E. (2005). Saying no to negativity: The effects of context and motivation to control prejudice on automatic evaluative responses. *Journal of Experimental Social Psychology*, 41(1), 19–35.  
<http://doi.org/10.1016/j.jesp.2004.05.002>
- Mann, T. C., Katz, J., Ferguson, M. J., & Goncalo, J. (in preparation). Implicitly creative: The rapid formation of implicit trait impressions beyond positivity and negativity. Manuscript in preparation.
- Marien, H., Custers, R., Hassin, R. R., & Aarts, H. (2012). Unconscious goal activation and the hijacking of the executive function. *Journal of Personality and Social Psychology*, 103, 399-415. <http://doi.org/10.1037/a0028955>
- McConnell, A. R., & Rydell, R. J. (2014). The systems of evaluation model: A dual-systems approach to attitudes. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 204-218). New York: Guilford.
- McConnell, A. R., Rydell, R. J., Strain, L. M., & Mackie, D. M. (2008). Forming implicit and explicit attitudes toward individuals: Social group association cues. *Journal of Personality and Social Psychology*, 94(5), 792–807.
- Mendoza, S. A., Gollwitzer, P. M., & Amodio, D. M. (2010). Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions. *Personality and Social Psychology Bulletin*, 36(4), 512–523.  
<http://doi.org/10.1177/0146167210362789>



- Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, 36, 630 – 633.
- Monroe, B. M., & Read, S. J. (2008). A general connectionist model of attitude structure and change: The ACS (Attitudes as Constraint Satisfaction) model. *Psychological Review*, 115(3), 733–759. <http://doi.org/10.1037/0033-295X.115.3.733>
- Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, 7(3), 134–140. [http://doi.org/10.1016/S1364-6613\(03\)00028-7](http://doi.org/10.1016/S1364-6613(03)00028-7)
- Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, 132(2), 297–326. <http://doi.org/10.1037/0033-2909.132.2.297>
- Morsella, E. (2005). The Function of phenomenal states: Supramodular interaction theory. *Psychological Review*, 112(4), 1000–1021.
- Moskowitz, G. B., Gollwitzer, P. M., Wasel, W., & Schaal, B. (1999). Preconscious control of stereotype activation through chronic egalitarian goals. *Journal of Personality and Social Psychology*, 77, 167–184.
- Moskowitz, G. B., & Ignarri, C. (2009). Implicit volition and stereotype control. *European Review of Social Psychology*, 20, 97–145.
- Moskowitz, G. B., & Li, P. (2011). Egalitarian goals trigger stereotype inhibition: A proactive form of stereotype control. *Journal of Experimental Social Psychology*, 47(1), 103–116.
- Moskowitz, G. B., Salomon, A. R., & Taylor, C. M. (2000). Preconsciously controlling stereotyping: Implicitly activated egalitarian goals prevent the activation of stereotypes. *Social Cognition*, 18(2), 151–177.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–259.

- Nosek, B. A. (2007). Implicit–explicit relations. *Current Directions in Psychological Science*, 16(2), 65–69.
- Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1), 101–115. <http://doi.org/10.1037//1089-2699.6.1.101>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716–aac4716. <http://doi.org/10.1126/science.aac4716>
- Palminteri, S., Wyart, V., & Koechlin, E. (in press). The Importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*.  
<http://doi.org/10.1016/j.tics.2017.03.011>
- Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, 81, 181-192.
- Payne, B. K., Brown-Iannuzzi, J., Burkley, M., Arbuckle, N. L., Cooley, E., Cameron, C. D., & Lundberg, K. B. (2013). Intention invention and the affect misattribution procedure: Reply to Bar-Anan and Nosek (2012). *Personality and Social Psychology Bulletin*, 39(3), 375–386.
- Payne, B. K., Burkley, M., & Stokes, M. B. (2008). Why do implicit and explicit evaluation tests diverge? The role of structural fit. *Journal of Personality and Social Psychology*, 94, 16-31.
- Payne, B. K., Cheng, C. M., Govorun, O., & Stewart, B. (2005). An inkblot for evaluations: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89, 277-293.

- Payne, B. K., Hall, D. L., Cameron, C. D., & Bishara, A. J. (2010). A process model of affect misattribution. *Personality and Social Psychology Bulletin*, 36(10), 1397–1408.  
<http://doi.org/10.1177/0146167210383440>
- Payne, B. K., Lambert, A. J., & Jacoby, L. L. (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in race-based misperceptions of weapons. *Journal of Experimental Social Psychology*, 38(4), 384–396.
- Payne, K., & Lundberg, K. (2014). The affect misattribution procedure: Ten years of evidence on reliability, validity, and mechanisms. *Social and Personality Psychology Compass*, 8(12), 672–686. <http://doi.org/10.1111/spc3.12148>
- Pessiglione, M., Schmidt, L., Draganski, B., Kalisch, R., Lau, H., Dolan, R. J., & Frith, C. D. (2007). How the brain translates money into force: a neuroimaging study of subliminal motivation. *Science*, 316(5826), 904–906.
- Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin*, 37(4), 557–569. <http://doi.org/10.1177/0146167211400423>
- Petty, R. E., Briñol, P., & DeMarree, K. G. (2007). The meta-cognitive model (MCM) of evaluations: Implications for evaluation measurement, change, and strength. *Social Cognition*, 25, 657-686.
- Petty, R. E., Tormala, Z. L., Briñol, P., & Jarvis, W. B. G. (2006). Implicit ambivalence from evaluation change: An exploration of the PAST model. *Journal of Personality and Social Psychology*, 90(1), 21–41.

- Queller, S., & Smith, E. R. (2002). Subtyping versus bookkeeping in stereotype learning and change: Connectionist simulations and empirical findings. *Journal of Personality and Social Psychology*, 82(3), 300–313. <http://doi.org/10.1037//0022-3514.82.3.300>
- Radiolab (2015, September 7). The rhino hunter. *Radiolab*. Retrieved from <http://www.radiolab.org/story/rhino-hunter/>
- Ranganath, K. A., & Nosek, B. A. (2008). Implicit attitude generalization occurs immediately; explicit attitude generalization takes time. *Psychological Science*, 19(3), 249–254. <http://doi.org/10.1111/j.1467-9280.2008.02076.x>
- Ratliff, K. A., & Nosek, B. A. (2011). Negativity and outgroup biases in attitude formation and transfer. *Personality and Social Psychology Bulletin*, 37(12), 1692–1703. <http://doi.org/10.1177/0146167211420168>
- Ratliff, K. A., Swinkels, B. A. P., Klerx, K., & Nosek, B. A. (2012). Does one bad apple(juice) spoil the bunch? Implicit attitudes toward one product transfer to other products by the same brand. *Psychology and Marketing*, 29(8), 531–540. <http://doi.org/10.1002/mar.20540>
- Richards, Z., & Hewstone, M. (2001). Subtyping and subgrouping: Processes for the prevention and promotion of stereotype change. *Personality and Social Psychology Review*, 5(1), 52–73.
- Rule, N. O., & Ambady, N. (2008). Brief exposures: Male sexual orientation is accurately perceived at 50-ms. *Journal of Experimental Social Psychology*, 44, 1100–1105. <http://dx.doi.org/10.1016/j.jesp.2007.12.001>

- Rule, N. O., Tskhay, K. O., Freeman, J. B., & Ambady, N. (2014). On the interactive influence of facial appearance and explicit knowledge in social categorization. *European Journal of Social Psychology, 44*(6), 529–535. <http://doi.org/10.1002/ejsp.2043>
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*(1), 145–172. <http://doi.org/10.1037/0033-295X.110.1.145>
- Rydell, R. J., & Gawronski, B. (2009). I like you, I like you not: Understanding the formation of context-dependent automatic attitudes. *Cognition & Emotion, 23*(6), 1118–1152.
- Rydell, R. J., & McConnell, A. R. (2006). Understanding implicit and explicit evaluation change: A systems of reasoning analysis. *Journal of Personality and Social Psychology, 91*(6), 995–1008.
- Rydell, R. J., McConnell, A. R., Mackie, D. M., & Strain, L. M. (2006). Of two minds: Forming and changing valence-inconsistent implicit and explicit evaluations. *Psychological Science, 17*(11), 954–958.
- Rydell, R. J., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2007). Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology, 37*(5), 867–878.
- Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioural cues in black and white children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology, 39*, 590–598.
- Sava, F. A., MaricuToiu, L. P., Rusu, S., Macsinga, I., Virgă, D., Cheng, C. M., & Payne, B. K. (2012). An Inkblot for the Implicit Assessment of Personality: The Semantic Misattribution Procedure. *European Journal of Personality, 26*(6), 613–628. <http://doi.org/10.1002/per.1861>

- Schiller, D., Kanen, J. W., LeDoux, J. E., Monfils, M. H., & Phelps, E. A. (2013). Extinction during reconsolidation of threat memory diminishes prefrontal cortex involvement. *Proceedings of the National Academy of Sciences*, *110*(50), 20040–20045.  
<http://doi.org/10.1073/pnas.1320322110>
- Schein, C., & Gray, K. (in press). Moralization and harmification: The dyadic loop explains how the innocuous becomes harmful and wrong. *Psychological Inquiry*.
- Sevenster, D., Beckers, T., & Kindt, M. (2012). Retrieval per se is not sufficient to trigger reconsolidation of human fear memory. *Neurobiology of Learning and Memory*, *97*(3), 338–345.
- Sevenster, D., Beckers, T., & Kindt, M. (2013). Prediction error governs pharmacologically induced amnesia for learned fear. *Science*, *339*(6121), 830–833.
- Sevenster, D., Beckers, T., & Kindt, M. (2014). Prediction error demarcates the transition from retrieval, to reconsolidation, to new learning. *Learning & Memory*, *21*(11), 580–584.  
<http://doi.org/10.1101/lm.035493.114>
- Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & Groom, C. J. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review*, *115*(2), 314–335.
- Shurick, A. A., Hamilton, J. R., Harris, L. T., Roy, A. K., Gross, J. J., & Phelps, E. A. (2012). Durable effects of cognitive restructuring on conditioned fear. *Emotion*, *12*(6), 1393–1397. <http://doi.org/10.1037/a0029143>
- Sinclair, S., Lowery, B. S., Hardin, C. D., & Colangelo, A. (2005). Social tuning of automatic racial attitudes: the role of affiliative motivation. *Journal of Personality and Social Psychology*, *89*(4), 583–592.

- Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, 105, 131–142.
- Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, 4(2), 108–131.
- Snyder, M., Tank, E. D., & Berscheid, E. (1977). Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes. *Journal of Personality and Social Psychology*, 35, 656–666.
- Spruyt, A., De Houwer, J., Hermans, D., & Eelen, P. (2007). Affective priming of nonaffective semantic categorization responses. *Experimental Psychology (Formerly "Zeitschrift Für Experimentelle Psychologie")*, 54(1), 44–53. <http://doi.org/10.1027/1618-3169.54.1.44>
- Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social Psychology*, 37(10), 1660–1672. <http://doi.org/10.1037/0022-3514.37.10.1660>
- Srull, T. K., & Wyer, R. S. (1980). Category accessibility and social perception: Some implications for the study of person memory and interpersonal judgments. *Journal of Personality and Social Psychology*, 38(6), 841–856.
- Steinfurth, E. C. K., Kanen, J. W., Raio, C. M., Clem, R. L., Haganir, R. L., & Phelps, E. A. (2014). Young and old Pavlovian fear memories can be modified with extinction training during reconsolidation in humans. *Learning & Memory*, 21(7), 338–341. <http://doi.org/10.1101/lm.033589.113>

- Stewart, B. D., & Payne, B. K. (2008). Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin*, 34(10), 1332–1345.
- Stewart, B. D., von Hippel, W., & Radvansky, G. A. (2009). Age, race, and implicit prejudice: Using process dissociation to separate the underlying components. *Psychological Science*, 20(2), 164–168.
- Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review*, 8(3), 220–247.
- Tabak, J. A., & Zayas, V. (2012). The roles of featural and configural face processing in snap judgments of sexual orientation. *PLoS One*, 7(5), 1–7, e3667. <http://doi.org/10.1371/journal.pone.0036671>
- Teige-Mocigemba, S., & Klauer, K. C. (2008). “Automatic” evaluation? Strategic effects on affective priming. *Journal of Experimental Social Psychology*, 44(5), 1414–1417. <http://doi.org/10.1016/j.jesp.2008.04.004>
- Teige-Mocigemba, S., & Klauer, K. C. (2013). On the controllability of evaluative-priming effects: Some limits that are none. *Cognition & Emotion*, 27(4), 632–657. <http://doi.org/10.1080/02699931.2012.732041>
- Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social Attributions from Faces: Determinants, Consequences, Accuracy, and Functional Significance. *Annual Review of Psychology*, 66(1), 519–545. <http://doi.org/10.1146/annurev-psych-113011-143831>



- Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: Evidence from a false recognition paradigm. *Journal of Personality and Social Psychology*, 83(5), 1051–1065. <http://doi.org/10.1037//0022-3514.83.5.1051>
- Todorov, A., & Uleman, J. S. (2003). The efficiency of binding spontaneous trait inferences to actors' faces. *Journal of Experimental Social Psychology*, 39(6), 549–562. [http://doi.org/10.1016/S0022-1031\(03\)00059-3](http://doi.org/10.1016/S0022-1031(03)00059-3)
- Todorov, A., & Uleman, J. S. (2004). The person reference process in spontaneous trait inferences. *Journal of Personality and Social Psychology*, 87(4), 482–493. <http://doi.org/10.1037/0022-3514.87.4.482>
- Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, 117(2), 440–463. <http://doi.org/10.1037/a0018963>
- Vallone, R. P., Griffin, D. W., Lin, S., & Ross, L. (1990). Overconfident prediction of future actions and outcomes by self and others. *Journal of Personality and Social Psychology*, 58, 582-592.
- Van Bavel, J. J., Jenny Xiao, Y., & Cunningham, W. A. (2012). Evaluation is a dynamic process: Moving beyond dual system models. *Social and Personality Psychology Compass*, 6(6), 438–454. <http://doi.org/10.1111/j.1751-9004.2012.00438.x>
- van Gaal, S., Naccache, L., Meuwese, J. D. I., van Loon, A. M., Leighton, A. H., Cohen, L., & Dehaene, S. (2014). Can the meaning of multiple words be integrated unconsciously? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1641), 20130212–20130212.
- Weber, R., & Crocker, J. (1983). Cognitive processes in the revision of stereotypic beliefs. *Journal of Personality and Social Psychology*, 45(5), 961-977.

- Wegner, D. M. (2003). The mind's best trick: how we experience conscious will. *Trends in Cognitive Sciences*, 7(2), 65–69. [http://doi.org/10.1016/S1364-6613\(03\)00002-0](http://doi.org/10.1016/S1364-6613(03)00002-0)
- Wegner, D. M., Sparrow, B., & Winerman, L. (2004). Vicarious agency: Experiencing control over the movements of others. *Journal of Personality and Social Psychology*, 86(6), 838–848. <http://doi.org/10.1037/0022-3514.86.6.838>
- Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist*, 54(7), 480–392.
- Wells, B. M., Skowronski, J. J., Crawford, M. T., Scherer, C. R., & Carlston, D. E. (2011). Inference making and linking both require thinking: Spontaneous trait inference and spontaneous trait transference both rely on working memory capacity. *Journal of Experimental Social Psychology*, 47(6), 1116–1126. <http://doi.org/10.1016/j.jesp.2011.05.013>
- Whalen, P. J., Rauch, S. L., Etkoff, N. L., McInerney, S. C., Lee, M. B., & Jenike, M. A. (1998). Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *The Journal of Neuroscience*, 18(1), 411–418.
- Wickelmaier, F. (2011). Multinomial processing tree models in R. Presented at the R User Conference 2011, August 16–18, Coventry, UK.
- Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after 100 ms exposure to a face. *Psychological Science*, 17, 592–598.
- Wilson, T. D. (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, MA: Harvard University Press.

- Wilson, T. D., Dunn, D. S., Kraft, D., & Lisle, D. J. (1989). Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123- 205). Orlando, FL: Academic Press.
- Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review*, 107(1), 101–126. <http://doi.org/10.1037/0033-295X.107.1.101>
- Wittenbrink, B., Judd, C. M., & Park, B. (1997). Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures. *Journal of Personality and Social Psychology*, 72(2), 262–274.
- Wojnowicz, M. T., Ferguson, M. J., Dale, R., & Spivey, M. J. (2009). The self-organization of explicit attitudes. *Psychological Science*, 20(11), 1428–1435.  
<http://doi.org/10.1111/j.1467-9280.2009.02448.x>
- Wyer, N. A. (2010). You never get a second chance to make a first (implicit) impression: The role of elaboration in the formation and revision of implicit impressions. *Social Cognition*, 28(1), 1–19.
- Wyer, N. A. (2016). Easier done than undone... by some of the people, some of the time: The role of elaboration in explicit and implicit group preferences. *Journal of Experimental Social Psychology*, 63, 77–85. <http://doi.org/10.1016/j.jesp.2015.12.006>
- Yogeeswaran, K., & Dasgupta, N. (2014). The devil is in the details: Abstract versus concrete construals of multiculturalism differentially impact intergroup relations. *Journal of Personality and Social Psychology*, 106(5), 772–789. <http://doi.org/10.1037/a0035830>

Zayas, V., & Shoda, Y. (2015). Love you? Hate you? Maybe it's both: Evidence that significant others trigger bivalent-priming. *Social Psychological and Personality Science*, 6(1), 56–64. <http://doi.org/10.1177/1948550614541297>

## Appendix: Immersive Materials Condition of Study 12

### Instructions Page

On the following pages, we will ask you to read statements describing fictional events, and would like you to imagine that these events are actually occurring, and happening to YOU.

It is very important that you really try to *immerse yourself* in the situation being described.

Please read the information carefully, as you will be asked questions on it later.

### Materials (1 paragraph per screen)

Imagine you are driving home one day from a long day of work, and you get a searing pain in your side. The pain is so bad that you have to immediately pull over. It feels like someone has just stabbed you in your abdomen, and you double over while holding your side.

You start to breath heavily, and grasp for your phone. In a panic, you quickly call your best friend, who advises you to call an ambulance right away. The pain is wave-like, and sometimes dull and deep, and sometimes intense and unbearable.

You agree to call the ambulance, which arrives in about 6 minutes. The medics carefully put you into the ambulance, and race to the hospital with the sirens blaring.

While in the ambulance, they take your vitals. You have a high fever, and your blood pressure is increasing. You are sweating and dizzy, and are very afraid. They rule out any problem with your appendix, and call the staff at the hospital to advise them that they are bringing you in.

Once you get to the hospital, you meet your nurse, [Jonathan/Elizabeth], and lead doctor, [Elizabeth/Jonathan]. [Images of Jonathan and Elizabeth are first displayed on this page.]

They read the chart from the ambulance medics, and rush you into an exam room to take further tests. [Jonathan/Elizabeth], the nurse, is taking your vitals again, and [Elizabeth/Jonathan], the doctor, is deciding which course to follow. You will need to ask them questions and so you should take a minute to remember their face and name carefully and their role in diagnosing your condition.